# Inhalt

methoden, daten, analysen · 2013, Jg. 7(2), S. 145-152 DOI: 10.12758/mda.2013.009

# Standardisierte Kurzskalen zur Erfassung psychologischer Merkmale in Umfragen

# Short Scales for the Assessment of Psychological Constructs in Surveys

## *Einleitung*

## *Introduction*

*Beatrice Rammstedt, Christoph J. Kemper, Jürgen Schupp*

*Zusammenfassung*
Psychologische Merkmale gewinnen zunehmend an Bedeutung in der sozialwissenschaftlichen Umfrageforschung. Da psychologische Instrumente vielfach einen individualdiagnostischen Entstehungshintergrund haben, sind sie meist viel zu umfangreich für einen Einsatz in Umfragen. Für solche Erhebungssituationen sind extrem kurze aber auch für die gesamte Bevölkerungsbreite validierte Verfahren angemessener. Im Rahmen dieses Sonderhefts werden sechs für diese Zwecke entwickelte und validierte Erhebungsinstrumente zur Erfassung von Merkmalen wie Persönlichkeit, Attraktivität oder Intelligenz vorgestellt. Sämtliche dieser Verfahren stehen der gesamten sozialwissenschaftlichen Profession zur Verfügung und können kostenfrei genutzt werden.

*Abstract*
Psychological constructs become more and more important in social surveys. Scales assessing these constructs are usually developed with the focus on an individual diagnosis. Therefore, they are in most cases much too lengthy to be assessed in surveys. For such settings extremely short measures validated for the total population are needed. In the present special issue the development and psychometric properties of six short scales are presented assessing constructs like personality, physical attractiveness and intelligence. All of the instruments presented can be used by the scientific community free of charge.

Weltweit sind seit einigen Jahren deutliche Bestrebungen zu beobachten, psychologische Merkmale wie die individuelle Persönlichkeit, Kontrollüberzeugungen oder Intelligenz in sozialwissenschaftlichen Umfragen zu erheben, um mittels dieser Merkmale sozialwissenschaftliche Fragestellungen besser erklären zu können. Der Zusammenhang zwischen psychologischen Merkmalen und sozioökonomischen Erfolgsgrößen gilt inzwischen als gut gesichert (Gottfredson 1997; Gottfredson/Deary 2004; Schmidt/Hunter 1998; Strenze 2007). So konnte gezeigt werden, dass die kognitiven Fähigkeiten einer Person, insbesondere deren Intelligenz, der beste Prädiktor für ein erfolgreiches Leben ist. Personen mit hoher kognitiver Leistungsfähigkeit sind vergleichsweise erfolgreicher in Schule, Studium, Ausbildung, Beruf und im Privatleben. So haben intelligentere Personen im Mittel nicht nur ein höheres Einkommen oder eine höhere Position im Beruf, sie lassen sich auch seltener scheiden und werden seltener delinquent oder arbeitslos. Auch viele weitere sozialwissenschaftlich relevante Prozesse und Phänomene, die mitunter weitreichende Implikationen für den Einzelnen, seine Mitmenschen oder die Gesellschaft als Ganzes haben, werden von psychologischen Merkmalen beeinflusst. So kann auf Grundlage von psychologischen Merkmalen beispielsweise das Wahlverhalten prädiziert werden (Schumann/Schoen 2005) aber auch, auf Basis sogenannter „noncognitive skills", ökonomischer Erfolg im Lebensverlauf besser erklärt werden (Almlund et al. 2012). Auch die individuelle Gesundheit und sogar die Mortalität sind durch psychologische Merkmale beeinflusst (Allison/Guichard/Fung/Gilain 2003; Arthur/Graziano 1996; Rasmussen/Scheier/Greenhouse 2009).

Vor dem Hintergrund der durch derartige Befunde belegten Nützlichkeit psychologischer Merkmale zur Verbesserung von Deskription und Prädiktion wissenschaftlich und gesellschaftlich relevanter Prozesse und Phänomene, forderte bereits vor einigen Jahren der Ökonomie-Nobelpreisträgers James Heckman, dass sozialwissenschaftliche Studien neben Intelligenztests auch validierte Persönlichkeitsskalen beinhalten sollten (Borghans/Ducksworth/Heckman/ter Weel 2008). Dieser Standpunkt wird auch von anderen Forscherinnen und Forschern (Goldberg 2005; Rammstedt 2010a) und Institutionen (Rat für Sozial- und Wirtschaftsdaten 2010) vertreten. Seit einigen Jahren sind diesbezüglich deutliche Bestrebungen erkennbar. Im Sozio-oekonomischen Panel (SOEP) wurden bereits in den 90er Jahren Kontrollüberzeugungen erhoben. So wurde 2004 erstmals Risikoaversion, in 2005 Persönlichkeit und Reziprozität (Schupp/Spieß/Wagner 2008) und in 2006 auch grundlegende Intelligenzmaße (Lohmann/Spieß/Groh-Samberg/Schupp 2009) erfasst. Darüber hinaus wurden Skalen zur Erfassung psychologischer Merkmale in weitere quer- wie längsschnittliche Untersuchungen aufgenommen, zum Beispiel in das International Social Survey Programme (ISSP), das Household, Income

and Labour Dynamics in Australien (HILDA), die UK Household Longitudinal Study (UKHLS) und den DNB Household Survey (DHS). Auch in der derzeit begonnenen umfangreichsten Längsschnittstudie Deutschlands auf Basis von 200.000 Erwachsenen, der Nationalen Kohorte, werden psychologische Kurzskalen ins Befragungsprogramm integriert. Der Bedarf an Erhebungsinstrumenten zur Messung psychologischer Merkmale in Large-scale-Studien ist demnach gegeben und wird in den kommenden Jahren vermutlich steigen.

Trotz des zunehmenden Interesses an psychologischen Merkmalen ist gleichwohl ein Mangel an geeigneten Erhebungsverfahren zu konstatieren. Zwar existieren für die meisten dieser Merkmale psychometrisch geprüfte standardisierte Erhebungsverfahren, diese sind jedoch in der Regel für den Einsatz in der psychologischen Individualdiagnostik entwickelt worden und daher deutlich zu umfangreich für den Einsatz in repräsentativen Large-scale-Studien mit in der Regel sehr heterogenen Bevölkerungsgruppen. Die meisten der psychologischen Standardverfahren sind für einen Einsatz in derartigen Erhebungen schlicht zu zeitaufwendig und vielfach auch für Befragte aufgrund ihrer Länge ermüdend. So umfasst das etablierteste Verfahren zur Erfassung der fünf Persönlichkeitsdimensionen, das NEO-PI-R, 240 Items und benötigt in einer Interviewsituation oder im Falle von Selbstausfüllen des Fragebogens eine Bearbeitungszeit von deutlich über einer halben Stunde. Das im deutschen Sprachraum gängigste Verfahren zur Messung internaler und externaler Kontrollüberzeugung, der sogenannte Fragebogen zur Kompetenz- und Kontrollüberzeugung (FKK; Krampen 1991), umfasst immerhin noch 32 Items und ist somit ebenfalls nicht unter fünf Minuten bearbeitbar. Im Bereich der Intelligenzmessung sind die Bearbeitungszeiten entsprechender Testverfahren sogar noch deutlich länger. Diese Beispiele zeigen, dass zwar für zahlreiche psychologische Merkmale erprobte Erhebungsinstrumente existieren, diese aber aufgrund ihres Umfangs nicht für den Einsatz in Untersuchungskontexten mit extremen zeitlichen Limitationen, wie der Umfrageforschung, die in der Regel auf die freiwillige und unentgeltliche Teilnahmebereitschaft der Respondenten angewiesen ist, geeignet sind.

In Ermangelung angemessener Erhebungsverfahren für psychologische Merkmale mit hoher Relevanz für die sozialwissenschaftliche Umfrageforschung, werden von den Umfrageforscherinnen und -forschern bislang vielfach Ad-hoc-Instrumente entwickelt. Dies kann zu zwei Problemen führen. Zum einen werden die Verfahren in der Regel nicht hinreichend hinsichtlich ihrer Reliabilität und Validität geprüft, sodass deren psychometrische Güte teils unbekannt ist. Zum anderen erfolgen Erhebungen derselben Merkmale mit teilweise unterschiedlichen Messinstrumenten. So wurden beispielsweise im ISSP, im SOEP, im UK Householdpanel und

im HILDA die fünf grundlegenden Persönlichkeitsdimensionen Extraversion, Verträglichkeit, Gewissenhaftigkeit, Emotionale Stabilität und Offenheit für Erfahrungen erfasst, jedoch in jeder Umfrage ein aus unterschiedlichen Items bestehendes Erhebungsinstrument eingesetzt. Darüber hinaus wurden die meisten dieser Erhebungsinstrumente entweder lediglich basierend auf Pretestdaten psychometrisch überprüft oder ad-hoc in Anlehnung an entsprechende psychologische Konstrukte für die jeweilige Studie gänzlich neu entwickelt, sodass bislang keine psychometrischen Kennwerte vorliegen. Bei anderen psychologischen Merkmalen sieht die Lage vergleichbar aus.

Um diesem Problem zu begegnen und künftig die Datenqualität sozialwissenschaftlicher Umfragen zu verbessern, wurden u.a. von den Autoren dieses Sonderhefts diverse Erhebungsinstrumente für psychologische Merkmale entwickelt und umfassend validiert. Zentrales Kriterium dieser Instrumente ist, dass sie nicht nur reliabel und valide sind, sondern auch möglichst kurz und leicht verständlich, damit sie auch in Erhebungssituationen mit starken zeitlichen Limitationen und bei unterschiedlichen Bevölkerungsgruppen eingesetzt werden können. Jedes der vorgestellten Verfahren wurde dafür in einem mehrstufigen Prozess entwickelt und auf Basis umfangreicher heterogener Stichproben validiert.

Ziel dieses Sonderhefts der Zeitschrift Methoden Daten Analysen ist es, Forscherinnen und Forschern exemplarisch sechs dieser kürzlich entwickelten ökonomischen Erhebungsinstrumente vorzustellen. Allen sechs Verfahren ist gemein, dass sie sozialwissenschaftlich relevante psychologische Merkmale erfassen.

In den ersten beiden Beiträgen des Sonderhefts werden Entwicklung und Validierung von Kurztests zur Erfassung von Intelligenz beschrieben. Intelligenz gilt, wie oben bereits beschrieben, als guter Prädiktor für eine Vielzahl gesellschaftlich relevanter Prozesse und Phänomene. Nach dem noch immer aktuellen Intelligenzmodell von Cattell (1971) kann Intelligenz grob in fluide Intelligenz (gf) und kristalline Intelligenz (gc) unterschieden werden. Kristalline Intelligenz spiegelt die Einflüsse von Lernen und Akkulturation wider und umfasst somit das gesamte Wissen, das Menschen im Laufe ihres Lebens erwerben und zum Problemlösen einsetzen. Sie soll sich laut Cattell (1971) in Testleistungen zeigen, die auf die Inhalte formaler Bildung abzielen. Im Gegensatz zur kristallinen Intelligenz spiegelt die fluide Intelligenz nach Cattell die dekontextualisierte Fähigkeit wider, sich neuen Problemen oder Situationen anzupassen, ohne dass im wesentlichen Ausmaß auf frühere Lernerfahrungen zurückgegriffen werden muss (Cattell 1971). Die beiden hier im Folgenden dargestellten kurzen Intelligenztests, die Kurzform zu kristalliner Intelligenz des Berliner Test zur Erfassung fluider und kristalliner Intelligenz (BEFKI GC-K, Schipolowski et al. 2013) und der Hagener Matrizen-Test in seiner Kurzform

(HMT-S, Heydasch/Haubrich/Renner 2013) ermöglichen eine Messung genau dieser beiden Intelligenzaspekte. Im BEFKI GC-K wird kristalline Intelligenz anhand von Antwortwahlaufgaben zu Wissen aus den Natur-, Geistes- und Sozialwissenschaften operationalisiert. Der HMT-S erfasst die Fähigkeit zum schlussfolgernden Denken anhand von Matrizenaufgaben, die etablierte Indikatoren für fluide Intelligenz sind (Carroll 1993). Mit diesen beiden Tests stehen der Umfrageforschung nun erstmals ökonomische und kostenfrei nutzbare Erhebungsinstrumente zur Messung von Intelligenz zur Verfügung. Die Auswahl eines der Tests sollte stets von der untersuchten Fragestellung abhängig gemacht werden. Stehen in einer Erhebung bspw. Korrelate der auf Lernen, Erfahrung und Wissen beruhenden Problemlösefähigkeit im Vordergrund, sollte der BEFKI GC-K eingesetzt werden. Stehen hingegen Korrelate der von Lernerfahrungen und Akkulturation weitgehend unabhängigen Problemlösefähigkeit im Fokus der Untersuchung, ist der Einsatz des HMT-S zu empfehlen.

Im dritten Beitrag von Lutz et al. wird das Attraktivitätsrating 1 (AR1), eine Ratingskala zur Messung von physischer Attraktivität vorgestellt. Aufgrund ihrer Rolle bei einer Vielzahl sozialer Phänomene und Prozesse ist die physische Attraktivität für die sozialwissenschaftliche Forschung und darüber hinaus, ein hochgradig relevantes Personenmerkmal. Zahlreiche Studien belegen, dass attraktive Personen im Vergleich zu weniger attraktiven in der sozialen Interaktion oft bevorzugt werden und Vorteile haben: Attraktivere bekommen häufiger gut bezahlte berufliche Positionen mit hohem Prestige (z.B. Schuler/Berger 1979; Umberson/Hughes 1987), attraktivere Kriminelle erhalten mildere Urteile als weniger Attraktive (Sigall/Ostrove 1975), Attraktiveren wird mehr Aufmerksamkeit geschenkt und ihnen wird häufiger geholfen (Langlois et al. 2000). Das AR1 erlaubt eine effiziente Erfassung dieses Personenmerkmals mit einem Item und wendet außerdem ein innovatives Messkonzept an, das den in der Fachliteratur berichteten Schwächen bei der üblichen Messmethode Rechnung trägt. Hierbei wird die physische Attraktivität einer Zielperson, z.B. einer Befragungsperson, von einem Beurteiler, z.B. einem Interviewer, relativ zu einem Vergleichsmaßstab eingeschätzt. Als Vergleichsmaßstab wird das Bild einer durchschnittlich attraktiven Person desselben Geschlechts und eines ähnlichen Alters wie die Zielperson verwendet. Somit werden durch Merkmale des Beurteilers, wie dessen Alter und Geschlecht, verursachte Verzerrungen von Attraktivitätseinschätzungen vermieden.

In dem vierten Beitrag wird das zehn Items umfassende Big Five Inventory (BFI-10, Rammstedt et al. 2013), ein bereits recht etabliertes Verfahren zur ökonomischen Erfassung der oben genannten fünf zentralen Persönlichkeitsdimensionen, der sogenannten Big Five, dargestellt. Die Big Five haben sich in den letzten Jah-

ren als erfolgreiche Prädiktoren für verschiedene individuelle wie gesellschaftliche Aspekte erwiesen. So leben emotional stabilere und gewissenhaftere Personen durchschnittlich gesünder und deutlich länger (vgl. Roberts/Kuncel/Shiner/Caspi/Goldberg 2007). Auch die Berufswahl, der berufliche Erfolg oder politisches Wahlverhalten werden von der individuellen Persönlichkeit geprägt. Die Erfassung dieser Big Five-Persönlichkeitsdimensionen in sozialwissenschaftlichen Large-scale-Surveys ist so weit verbreitet, dass sie inzwischen in die meisten Studien im deutschen Sprachraum einbezogen werden. Allerdings – wie oben dargestellt – erfolgt diese Erfassung mit sehr unterschiedlichen Maßen. Der Beitrag stellt eine aktuelle Validierung des BFI-10 anhand einer umfangreichen, bevölkerungsrepräsentativen Stichprobe dar.

Der fünfte Beitrag (Beierlein et al. 2013a) stellt ein Drei-Item-Inventar (ASKU) zur Erfassung der allgemeinen Selbstwirksamkeit, definiert als die Einschätzung der eigenen Kompetenz zur Erreichung eines Ziels, dar. Da Selbstwirksamkeitserwartungen mit Aspekten der Arbeit, der Gesundheit und sozialer Beziehungen in Verbindung stehen, sind sie für die interdisziplinäre Surveyforschung von hohem Interesse. Allerdings hat es bislang an einem surveykompatiblen, d.h. ökonomischen Verfahren zur Erfassung des Merkmals gemangelt – einem Mangel dem mit der vorliegenden Entwicklung Rechnung getragen wurde.

Der sechste und letzte Beitrag (Beierlein et al. 2013b) adressiert das Merkmal der Ungerechtigkeitssensibilität, nämlich, wie leicht Ungerechtigkeit wahrgenommen und wie stark darauf reagiert wird. Unterschieden werden hierbei vier Perspektiven aus denen Ungerechtigkeit wahrgenommen werden kann: die Opfer-, die Beobachter-, die Nutznießer- und die Täterperspektive (Schmitt/Baumert/Fetchenhauer/Gollwitzer/Rothmund/Schlösser 2009). Das Merkmal der Ungerechtigkeitssensibilität ist insbesondere vor dem Hintergrund der aktuell in der Soziologie, der Politologie und der Ökonomie wieder aufblühenden Gerechtigkeitsforschung (Fetchenhauer/Goldschmidt/Hradil/Liebig 2010) von hohem Interesse für die sozialwissenschaftliche Profession. Studien konnten beispielsweise zeigen, dass Ungerechtigkeit in Zusammenhang mit sozialem Verhalten oder beruflichen und politischen Einstellungen steht. Aufbauend auf den vier Perspektiven der Ungerechtigkeitswahrnehmung, entwickelten Baumert, Beierlein, Schmitt, Kemper, Kovaleva, Liebig und Rammstedt (in Druck) vier, je zwei Item umfassende, Kurzskalen der etablierten Ungerechtigkeitssensibilitätsskala von Schmitt, Baumert, Gollwitzer und Maes (2010), deren Konstruktion und Validierung im sechsten Beitrag dargestellt wird.

# Literatur

Almlund, M., J. Heckman, A. L. Ducksworth und T. Kautz, 2011: Personal Psychology and Economics. S. 1-181 in: E. A. Hanushek, S. Machin und L.Wössman (Eds), Handbook of the Economics of Education. Amsterdam: Elsevier. Borghans, L., A. L. Ducksworth, J. J. Heckman und B. T. Weel, 2008: The Economics and Psychology of Personal Traits. The Journal of Human Resources 43(4): 972-1059.

Lohmann, H., C. K. Spieß, O. Groh-Samberg und J. Schupp, 2009: Analysepotenziale des Sozio-oekonomischen Panels (SOEP) für die empirische Bildungsforschung. Zeitschrift für Erziehungswissenschaft 12(2): 252-280.

Schupp, J., C. K. Spieß und G. G. Wagner, 2008: Die verhaltenswissenschaftliche Weiterentwicklung des Erhebungsprogramms des SOEP. Vierteljahrshefte zur Wirtschaftsforschung 77(3): 63-76.

Wichmann, H. E., R. Kaaks, W. Hoffmann, K.-H. Jöckel, K.-H. Greiser und J. Linseisen (2012): Die Nationale Kohorte. Bundesgesundheitsblatt 55(6-7): 781-789.

Baumert, A., C. Beierlein, M. Schmitt, C. J. Kemper, A. Kovaleva, Liebig, S. und B. Rammstedt, in Druck: Measuring four facets of Justice Sensitivity with two items each. Journal of Personality Assessment.

Carroll, J. B., 1993: Human cognitive abilities: A survey of factor-analytic studies. New York: Cambridge University Press.

Cattell, R. B., 1971: Abilities: Their Structure, Growth, and Action. Boston: Houghton Mifflin.Fetchenhauer, D., N. Goldschmidt, S. Hradil u. S. Liebig, 2010: Warum ist Gerechtigkeit wichtig? Antworten der empirischen Gerechtigkeitsforschung. München: Roman Herzog Institut.

Langlois, J. H., L. Kalakanis, A. J. Rubenstein, A. Larson, M. Hallam und M. Smoot, 2000: Maxims of myths of beauty? A meta-analytic and theoretical review. Psychological Bulletin 126: 390-423.

Roberts, B. W., N. R. Kuncel, R. Shiner, A. Caspi und L. R. Goldberg, 2007: The power of personality: The comparative validity of personality traits, socioeconomic status and cognitive ability for predicting important life outcomes. Perspectives on Psychological Science 2: 313-345.

Schmitt, M., A. Baumert, D. Fetchenhauer, M. Gollwitzer, T. Rothmund und T. Schlösser, 2009: Sensibilität für Ungerechtigkeit. Psychologische Rundschau 60: 8-22.

Schmitt, M., A. Baumert, M. Gollwitzer und J. Maes, 2010: The Justice Sensitivity Inventory: Factorial validity, location in the personality facet space, demographic pattern, and normative data. Social Justice Research 23: 211-238.

Schuler, H. und W. Berger, 1979: Physische Attraktivität als Determinante von Beurteilung und Einstellungsempfehlung. Psychologie und Praxis 23: 59-70.

Sigall, H. und N.Ostrove, 1975: Beautiful but dangerous: Effects of offender attractiveness and nature of the crime on juridic judgment. Journal of Personality and Social Psychology 31: 410-414.

Umberson, D. und M. Hughes, 1987: The impact of physical attractiveness on achievement and psychological well-being. Social Psychology Quarterly 50: 227-236. Schoen, H. und S. Schumann, 2005: Persönlichkeit: Eine vergessene Größe der empirischen Sozialforschung. Wiesbaden: Verlag für Sozialwissenschaften.

Lohmann, H., C. K. Spieß, O. Groh-Samberg und J. Schupp, 2009: Analysepotenziale des Sozio-oekonomischen Panels (SOEP) für die empirische Bildungsforschung. Zeitschrift für Erziehungswissenschaften 12 (2): 252-280.

Krampen, G., 1991: FKK – Fragebogen zu Kompetenz- und Kontrollüberzeugungen. Göttingen: Hogrefe.
Heydasch, T., J. Haubrich, und K. Renner, in Druck: Die Kurzform des Hagener Matrizen-Tests (HMT-S). Ein 6-Item Intelligenztest zum schlussfolgernden Denken.

| | |
|---|---|
| Anschrift der Autorin | Beatrice Rammstedt<br>GESIS – Leibniz-Institut für Sozialwissenschaften<br>Postfach 12 21 55<br>68072 Mannheim<br>E-Mail: beatrice.rammstedt@gesis.org |
| Ko-Autoren | Christoph J. Kemper<br>Institut für medizinische und pharmazeutische Prüfungsfragen (IMPP), Mainz |
| | Jürgen Schupp<br>Sozio-oekonomisches Panel (SOEP) am Deutschen Institut für Wirtschaftsforschung (DIW), Berlin |

methoden, daten, analysen · 2013, Jg. 7(2), S. 153-181 DOI: 10.12758/mda.2013.010

# BEFKI GC-K

## *Eine Kurzskala zur Messung kristalliner Intelligenz*

# BEFKI GC-K

## *A Short Scale for the Measurement of Crystallized Intelligence*

*Stefan Schipolowski, Oliver Wilhelm, Ulrich Schroeders, Anastassiya Kovaleva, Christoph J. Kemper und Beatrice Rammstedt*

*Zusammenfassung*

In aktuellen Intelligenzstrukturmodellen gehört kristalline Intelligenz ($g_c$) zu den am besten etablierten Fähigkeitsfaktoren. Dabei spiegelt $g_c$ die Einflüsse von Lernen und Akkulturation wider und umfasst somit alles Wissen, das Menschen im Laufe ihres Lebens erwerben und zum Problemlösen einsetzen. In diesem Beitrag beschreiben wir die Entwicklung einer Kurzskala zur Messung kristalliner Intelligenz mit fünfminütiger Bearbeitungszeit, die auf deklarativen Wissensfragen aus den Natur-, Geistes- und Sozialwissenschaften beruht. Aus einem umfangreichen Itempool wurde ein 32 Fragen umfassender Wissenstest zusammengestellt und einer bundesweit repräsentativen Stichprobe von 1.134 Erwachsenen vorgelegt. Anhand psychometrischer Kennwerte und der Beziehungen zu Kovariaten erfolgte eine Auswahl von 12 Items für die Kurzskala. Ein eindimensionales Messmodell für diese Itemauswahl wies eine gute Passung und eine hohe Reliabilität des latenten Faktors auf. In der Zielpopulation der erwachsenen deutschen Bevölkerung wurden keine substanziellen Boden- oder Deckeneffekte

*Abstract*

Crystallized intelligence ($g_c$) is a well-established cognitive ability factor that has been conceptualized as reflecting influences of learning, education, and acculturation. In this article, we describe the development of a short knowledge scale for the measurement of $g_c$ in five minutes administration time using declarative knowledge items from the sciences, the humanities, and civics. Based on a large item pool we compiled a 32-item knowledge test that was subsequently presented to a nationally representative sample of 1,134 German adults. In the next step, this data were used to derive a short 12-item knowledge scale. A unidimensional measurement model had satisfactory model fit and showed high reliability of the latent factor. There were no substantial floor or ceiling effects in the adult German population. Similar to the full scale, the short scale correlated highly positively with education (ISCED-97) and socio-economic status (ISEI) and was meaningfully related to self-reported knowledge and the Big Five personality traits. Therefore, the short knowledge scale allows for

beobachtet. Übereinstimmend mit der Langversion zeigten sich für die Kurzskala hohe Beziehungen zum Bildungsabschluss (ISCED-97) und sozioökonomischen Status (ISEI) sowie erwartungskonforme Korrelationen mit selbstberichtetem Wissen und den fünf Hauptdimensionen der Persönlichkeit (Big Five). Die Kurzskala ermöglicht folglich eine effiziente, reliable und valide Erfassung kristalliner Intelligenz im Rahmen der Umfrageforschung.

an efficient and valid measurement of crystallized intelligence in survey research.

# 1 Einleitung[1]

Die Messung psychologischer Merkmale wie kognitiver Fähigkeiten und Persönlichkeitseigenschaften gewinnt in den Sozial- und Wirtschaftswissenschaften zunehmend an Bedeutung (Rat für Sozial- und Wirtschaftsdaten 2010). So verweisen etwa Grabner und Stern (2010) auf das komplexe, bisher jedoch nicht hinreichend erforschte Zusammenspiel zwischen individuellen kognitiven Ressourcen und sozioökonomischen Variablen. Auf Personenseite kommt hier neben der dekontextualisierten Fähigkeit zum schlussfolgernden Denken bzw. dem Arbeitsgedächtnis insbesondere jenen Fähigkeiten eine zentrale Rolle zu, die auf Lernen, Erfahrung und Wissen beruhen und unter dem Begriff der kristallinen Intelligenz zusammengefasst werden. Wiederholt wurde jedoch darauf hingewiesen, dass bislang nur wenige für die Umfrageforschung geeignete, d.h. hinreichend kurze und effiziente Messinstrumente zur Erfassung psychologischer Merkmale vorliegen (Rammstedt/Kemper/Klein/Beierlein/Kovaleva 2012; Lang/Weiss/Stocker/von Rosenbladt 2007).

Der vorliegende Beitrag beschreibt die Entwicklung und Eigenschaften einer Kurzskala zur Erfassung kristalliner Intelligenz anhand deklarativer Wissensfragen in fünf Minuten Bearbeitungszeit. Die Datenbasis hierfür ist eine für die erwachsene Wohnbevölkerung Deutschlands repräsentative Stichprobe, die dementsprechend eine hohe Heterogenität bzgl. Alter und Bildung aufweist. Zuerst wird unter Rückgriff auf einflussreiche Intelligenztheorien das Konstrukt der kristallinen Intelligenz näher erläutert. Mit Blick auf die Validierung der Kurzskala folgt eine Darstellung zentraler Befunde zu den Beziehungen zwischen kristalliner Intelligenz, relevanten Personen- und Umweltmerkmalen sowie anderen psychologischen Konstrukten.

1 Während der Erstellung dieses Beitrags war Stefan Schipolowski Fellow der International Max Planck Research School „The Life Course: Evolutionary and Ontogenetic Dynamics (LIFE)".

## 1.1 Kristalline Intelligenz in konsensualen Intelligenzstrukturtheorien

Wissen und sprachliche Fähigkeiten spielen seit Beginn der psychometrischen Intelligenzforschung im späten 19. Jahrhundert eine bedeutende Rolle in Intelligenztheorien. So schlug Hermann Ebbinghaus bereits 1897 ein Verfahren zur „Prüfung geistiger Fähigkeiten" bei Schulkindern vor, das darauf beruhte, in kurzen Texten unvollständige Wörter sinnhaft zu vervollständigen und charakterisierte diese Anforderung explizit als „Intelligenzthätigkeit" (Ebbinghaus 1897: 414). Für Charles Spearman (1938) hingegen stellte non-verbales schlussfolgerndes Denken den eigentlichen Kern allgemeiner Intelligenz, des sog. g-Faktors, dar. Nach Binet und Simon (1905) sowie Hebb (1942) gehörte Cattell (1943) zu den ersten Intelligenzforschern, die diese beiden Intelligenzaspekte aufgriffen und gegenüberstellten. Statt eines einzigen Generalfaktors der Intelligenz postulierte Cattell zwei bedeutende Faktoren, die er als fluide Intelligenz ($g_f$) und kristalline Intelligenz ($g_c$) bezeichnete. Letztere zeigt sich nach Cattell (1971) in Leistungen, bei denen zuvor erlernte Fertigkeiten und Wissen die entscheidende Rolle spielen. Typische Indikatoren für $g_c$ bezeichnet Cattell als „schulische" oder „akademische" Tests, die auf die Inhalte formaler Bildung abzielen. Damit übereinstimmend wurden in den Studien in Cattells Labor (Cattell 1963; Horn/Cattell 1966; Cattell 1971) starke Ladungen von sprachnahen Aufgaben und Wissenstests auf dem Faktor $g_c$ berichtet (vgl. etwa Horn 1965). Aus Cattells theoretischen Schriften geht hervor, dass $g_c$ die Gesamtheit des Wissens umfasst, das Menschen im Laufe ihres Lebens erwerben und zum Problemlösen einsetzen. Aufgrund der mit steigendem Lebensalter zunehmenden Spezialisierung in Form unterschiedlicher Berufswahlen, verschiedenartiger Freizeitaktivitäten und Interessen würde eine umfassende Messung kristalliner Intelligenz im Erwachsenenalter demnach eine nahezu unendliche Vielfalt an Iteminhalten erfordern. Übereinstimmend betont auch Horn (1988) die Breite des $g_c$-Konstrukts und seine Nähe zum Generalfaktor der Intelligenz. In seiner Auflistung typischer Indikatoren spielen verbale Fähigkeiten eine prominente Rolle (z.B. „verbal knowledge") sowie Wissen in einer Vielzahl von Bereichen („information about the humanities, social and physical sciences, business and culture in general"; Horn 1988: 659). Horn und Noll (1997: 69) beschreiben kristalline Intelligenz dementsprechend als „akkulturiertes Wissen", das über Aufgaben gemessen wird, die „Tiefe und Breite des Wissens der dominanten Kultur" widerspiegeln.

Als weiteres einflussreiches Intelligenzmodell soll Carrolls (1993) Drei-Stratum-Theorie erwähnt werden, die nach wie vor als Status Quo der Intelligenzforschung gilt. Das Modell basiert auf der Reanalyse von mehr als 460 Datensätzen zu

kognitiven Fähigkeitskonstrukten und beschreibt verschiedene Schichten („Strata") mit Intelligenzfaktoren unterschiedlicher Breite. Zu den insgesamt acht Faktoren auf Stratum II gehören auch $g_f$ und $g_c$. Ähnlich wie Cattell betont auch Carroll mit Blick auf den $g_c$-Faktor die Rolle von Erfahrung, Lernen und Akkulturation, verschiebt jedoch die Definition von $g_c$ in Richtung sprachlicher Fähigkeiten wie Leseverstehen und Fremdsprachenkenntnissen (Carroll 1993: 626). Allerdings dokumentiert Carroll auch die hohen Ladungen von Wissenstests auf dem $g_c$-Faktor. Eine besondere Bedeutung kommt dem Stratum-I-Faktor „General Information" zu, der Unterschiede im Erwerb von Wissen jenseits von Sprachkenntnissen widerspiegelt (vgl. K0; Carroll 1993: 590 bzw. 634). Dieser Wissensfaktor gehört zu jenen Faktoren, die in Carrolls Reanalysen häufig die höchste oder zweithöchste Ladung auf einem übergeordneten $g_c$-Faktor aufwiesen (vgl. auch Carroll 2003). In Übereinstimmung mit den theoretischen Vorarbeiten Cattells untermauern diese Befunde die Position, dass eine Operationalisierung kristalliner Intelligenz im Erwachsenenalter Wissen aus möglichst vielen unterschiedlichen Bereichen berücksichtigen sollte.

## 1.2 Alters- und Geschlechtsunterschiede in kristalliner Intelligenz

Das Konzept der kristallinen Intelligenz wurde auch von Seiten der Entwicklungspsychologie über die Lebenspanne aufgegriffen (vgl. etwa Baltes 1987; Baltes/Staudinger/Lindenberger 1999), die wesentliche Erkenntnisse zum Entwicklungsverlauf von $g_c$ liefern konnte. Nach Baltes et al. (1999: 486ff) zeigt die kristalline Pragmatik der Kognition, die kulturell vermitteltes Wissen repräsentiert, einen deutlichen Anstieg im Kindes- und Jugendalter, bleibt im Verlauf des Erwachsenenalters weitgehend stabil und nimmt erst im sehr hohen Lebensalter ab. Auch Ackerman (2008) argumentiert in seinem Literaturüberblick, dass sich $g_c$ im Sinne von Allgemeinwissen und lexikalischem Wissen durch hohe Stabilität im Erwachsenenalter auszeichnet. Ein anderes Bild ergibt sich jedoch für spezialisiertes, bereichsspezifisches Wissen im Sinne von Expertise. So konnten Ackerman und Kollegen in mehreren Studien zeigen, dass im Erwachsenenalter teilweise deutliche Leistungszuwächse im bereichsspezifischen Wissen zu beobachten sind (Ackerman 2000; Ackerman/Rolfhus 1999; zusammenfassend Ackerman/Beier 2004). Diese Befunde illustrieren, dass die Beziehung kristalliner Intelligenz zum Alter auch von den gemessenen Inhalten abhängig ist. Wird $g_c$ über das im Kindes- und Jugendalter erworbene schulische Wissen gemessen, das unabhängig vom Lebensalter häufig abgerufen wird – wie es etwa für lexikalisches Wissen der Fall ist – sind kaum Veränderungen im Erwachsenalter zu beobachten; wird $g_c$ hingegen als Expertise operationalisiert,

die erst durch eine spezielle Ausbildung oder durch die Berufsausübung erworben wird, können auch im mittleren Erwachsenenalter substanzielle Zuwächse beobachtet werden.

Verschiedene Arbeiten aus der psychologischen Forschung haben sich mit Geschlechtsunterschieden in Wissensleistungen auseinandergesetzt. Hier sind insbesondere die Arbeiten von Lynn sowie aus der Arbeitsgruppe von Ackerman zu erwähnen. In Untersuchungen an Jugendlichen bzw. jungen Erwachsenen mit einer viele Wissensbereiche umfassenden Testbatterie wurde wiederholt ein Leistungsvorsprung im Allgemeinwissen zugunsten der Männer von etwa einer halben Standardabweichung gefunden ($d$ = 0.50 bis 0.60; Lynn/Irwing/Cammock 2001; Lynn/Irwing 2002; Lynn/Wilberg/Margraf-Stiksrud 2004). Ähnliche Ergebnisse berichten Ackerman, Bowen, Beier und Kanfer (2001), die eine Stichprobe von 320 Studierenden mit einer breit angelegten Wissenstestbatterie zu 19 spezifischen Wissensbereichen untersuchten. In 14 der 19 Wissensbereiche erzielten Männer signifikant bessere Ergebnisse, insbesondere in den naturwissenschaftlich-technischen und einigen sozialwissenschaftlichen Bereichen. Für keinen Wissensbereich ergaben sich signifikant bessere Resultate für Frauen; in vier Bereichen fanden sich keine oder sehr geringe Unterschiede. Insgesamt ergab sich für einen zusammengesetzten Gesamtwert zum deklarativen Wissen ein Unterschied von $d$ = 0.68 zugunsten der Männer. Einschränkend muss jedoch gesagt werden, dass diese geschlechtsbezogenen Disparitäten kulturspezifisch sind und nicht auf andere Kulturen übertragbar sein müssen.

## 1.3 Zusammenhänge kristalliner Intelligenz mit Bildung und Lernumwelt

Wie oben ausgeführt, ist die Abhängigkeit kristalliner Intelligenz von Lernen und Bildung das zentrale Definitionsmerkmal des Konstrukts. Aufgrund der Langfristigkeit des Wissenserwerbs und dessen Abhängigkeit von den zur Verfügung stehenden Lerngelegenheiten und -ressourcen sind folglich hohe Zusammenhänge zwischen $g_c$ und der Qualität und Quantität formaler Bildung zu erwarten, wie sie in formalen Bildungsabschlüssen und darauf beruhenden Indizes zum Ausdruck kommen (Cliffordson/Gustafsson 2008; Ceci 1991). Ebenso kann eine substanzielle positive Korrelation mit solchen Indikatoren und Indizes angenommen werden, die für den Wissenserwerb bedeutsame Ressourcen erfassen (Rowe/Jacobson/van den Oord 1999). Dies betrifft Maße des sozioökonomischen Status ebenso wie Fragen nach der Ausstattung des Haushalts, etwa zur Anzahl der verfügbaren Bücher (Ehmke/Siegle 2005).

## 1.4 Zusammenhänge kristalliner Intelligenz mit Selbsteinschätzungen des Wissens und Persönlichkeitskonstrukten

Neben einer Testung deklarativen Wissens über Wissensfragen mit mehreren Antwortalternativen, die eindeutig als richtig oder falsch zu werten sind, können auch Selbstberichte herangezogen werden. Derartige Selbsteinschätzungen des eigenen Wissens spiegeln jedoch neben der tatsächlichen Fähigkeitsausprägung weitere Varianzquellen wider (z.B. *faking-good*), so dass Selbsteinschätzungen des Wissens und objektive Wissenstests als unterschiedliche, wenngleich korrelierte Konstrukte aufzufassen sind (Furnham/Dissou 2007). Die Korrelationen zwischen entsprechenden Messungen sind somit zwar substanziell, fallen aber selbst unter Verwendung von Ansätzen zur Korrektur von Antwortverzerrungen deutlich niedriger aus als zwischen Messungen desselben Konstrukts (Hülür/Wilhelm/Schipolowski 2011; Rolfhus/Ackerman 1996; Paulhus/Harms 2004).

Eine Vielzahl an Studien hat sich mit den Zusammenhängen zwischen kognitiven Fähigkeiten einerseits und Persönlichkeitseigenschaften im Sinne typischen Verhaltens (Cronbach 1949) andererseits befasst. Hervorzuheben ist die Metaanalyse von Ackerman und Heggestad (1997), die anhand von 135 Studien die Beziehungen zwischen Persönlichkeits- und Fähigkeitskonstrukten beleuchtet. Die Autoren beschränken sich auf der Fähigkeitsseite nicht auf die Betrachtung allgemeiner Intelligenz, sondern differenzieren zwischen zehn verschiedenen Fähigkeitskonstrukten, darunter kristalline Intelligenz und eine zusammengefasste Kategorie Wissen/Achievement. Bei der Kategorisierung der gemessenen Konstrukte orientieren sich die Autoren an den Charakterisierungen der Fähigkeitsfaktoren bei Carroll (1993). Kristalline Intelligenz umfasst demnach sprachliche Leistungen und Allgemeinwissen (vgl. Carroll 1993: 599); zur Kategorie Wissen/Achievement zählen neben spezifischen Wissenstests auch Fachleistungstests etwa im Schulfach Biologie (vgl. Carroll 1993: 513). Die Persönlichkeitsskalen der Studien wurden anhand verschiedener in der Forschung etablierter Systeme klassifiziert, darunter das Modell der fünf Hauptdimensionen der Persönlichkeit (Big Five; für ausführliche Beschreibungen dieser Dimensionen siehe Ostendorf/Angleitner 2004). Die metaanalytische Betrachtung der korrelativen Zusammenhänge zwischen den Big Five-Faktoren und $g_c$ bzw. Wissen/Achievement ergab jeweils die mit Abstand höchsten Beziehungen für Offenheit. Hier wurde eine minderungskorrigierte Korrelation von .30 ($g_c$) bzw. .28 (Wissen/Achievement) ermittelt. Eine ebenfalls positive, aber mit .11 deutlich niedrigere Korrelation zeigte sich zwischen $g_c$ und Extraversion. Für Wissen/Achievement lag diese bei .05 und war nicht signifikant von null verschieden. Negative Zusammenhänge wurden hingegen mit Neurotizismus (Korrelationen

von -.09 mit $g_c$ und -.13 mit Wissen/Achievement) und Gewissenhaftigkeit gefunden, wobei die Korrelation zwischen Gewissenhaftigkeit und $g_c$ nicht signifikant war und für Wissen/Achievement mit -.19 zwar vom Betrag her höher ausfiel, diese Angabe jedoch nur auf einer einzigen Studie mit relativ geringer Fallzahl basiert. Die Korrelationen mit Verträglichkeit waren nicht signifikant von null verschieden.

Die vergleichsweise hohe Beziehung zwischen $g_c$ bzw. Wissen/Achievement und Offenheit ist sowohl empirisch gut belegt (von Stumm/Ackerman 2012) als auch aus theoretischer Perspektive plausibel. So zielen Items zur Messung von Offenheit unter anderem auf intellektuelle Neugier, das Bestreben, neues Wissen zu erwerben und auf kulturelles Engagement (Ostendorf/Angleitner 2004). Folgerichtig beschreibt Ackerman (1996) neben „typischem intellektuellem Engagement (TIE)" Offenheit als Persönlichkeitskonstrukt mit den höchsten Beziehungen zu $g_c$ und geisteswissenschaftlichem Wissen. Übereinstimmend argumentieren Ziegler, Danay, Heene, Asendorpf und Bühner (2012), dass Personen mit hohen Offenheitswerten mehr Zeit in Lernen und Wissenserwerb investieren.

## 1.5 Zusammenfassung und erwartete Zusammenhänge

Aus theoretischen und diagnostischen Überlegungen heraus sollte eine Operationalisierung von $g_c$ über deklaratives Wissen aus möglichst vielen verschiedenen Wissensbereichen erfolgen. Zu erwarten sind hohe positive Korrelationen zwischen kristalliner Intelligenz und der Qualität und Quantität formaler Bildung sowie mit Indikatoren, die für den Wissenserwerb bedeutsame Ressourcen erfassen, wie beispielsweise Maße des sozioökonomischen Status. Mit Blick auf Altersunterschiede ist in der Erwachsenenpopulation aufgrund der hohen Stabilität kristalliner Intelligenz von sehr geringen Effekten auszugehen, sofern die gemessenen Inhalte auf solche Wissensbestände abzielen, die in einer Vielzahl von Lernumwelten erworben werden können. In Anlehnung an die in der Literatur beschriebenen geschlechtsbezogenen Unterschiede in Wissensleistungen wird ein Vorsprung der Männer von etwa einer halben Standardabweichung bis zwei Drittel einer Standardabweichung erwartet. Des Weiteren ist von positiven Korrelationen von $g_c$ mit selbsteingeschätztem Wissen sowie mit dem Persönlichkeitsfaktor Offenheit auszugehen.

# 2 Methode

## 2.1 Stichprobe

Als Grundgesamtheit für die Stichprobenziehung wurde die Wohnbevölkerung der Bundesrepublik Deutschland im Alter von 18 Jahren und älter definiert. Es wurden auch Personen mit Zuwanderungshintergrund berücksichtigt, sofern sie die deutschsprachigen Fragen und Aufgaben verstehen und auf Deutsch antworten konnten. Die Stichprobe wurde mithilfe des ADM-Stichprobensystems F2F der Arbeitsgemeinschaft deutscher Marktforschungsinstitute gezogen. Dabei handelt es sich um ein komplexes mehrstufiges Ziehungsverfahren, in dem zunächst Flächen, dann Privathaushalte und schließlich Zielpersonen innerhalb der Haushalte nach einem Zufallsverfahren ausgewählt werden (für Details vgl. von der Heyde 2009). Nach diesem Verfahren wurde eine Stichprobe von 1206 Personen realisiert, die an der Erhebung teilnahmen. Im Anschluss wurden auf Basis des Zensus von GESIS Fallgewichte erstellt, um Repräsentativität für die o.g. Grundgesamtheit mit Blick auf Region (Ost- bzw. Westdeutschland), Geschlecht, Bildung und Alter zu gewährleisten. Grundlage der Gewichtung war ein reduzierter Datensatz von 1.134 Fällen nach Ausschluss unbrauchbarer Datenpunkte sowie von Personen ohne deutsche Staatsbürgerschaft, um für die Gewichtung eine eindeutige Definition der Grundgesamtheit zu ermöglichen. Die gewichtete Stichprobe umfasst somit 1.134 Erwachsene (52,2% weiblich) im Alter von 18 bis 93 Jahren ($M$ = 52 Jahre, $SD$ = 18 Jahre) aus dem gesamten Bundesgebiet.

## 2.2 Messinstrumente

In einem umfangreichen Fragebogen wurden verschiedene soziodemographische Merkmale der Teilnehmerinnen und Teilnehmer erfasst. Hierzu zählten Geburtsjahr und -monat, Geschlecht, Familienstand, Staatsangehörigkeit, erreichter bzw. (bei Schülerinnen und Schülern) angestrebter allgemeinbildender Schulabschluss, beruflicher Ausbildungsabschluss, Erwerbsstatus, berufliche Stellung und Haushaltsnettoeinkommen. Die Erfassung dieser Merkmale orientierte sich an den Demographischen Standards des Statistischen Bundesamtes (2010). Ergänzend wurde anhand von sieben Kategorien die Anzahl der Bücher im Elternhaus erfragt, das Geburtsland sowie die berufliche Stellung der Eltern, als der/die Teilnehmende 15 Jahre alt war. Anhand zweier fünfstufiger Skalen wurden vom Interviewer die Deutschkenntnisse und die soziale Schichtzugehörigkeit der Teilnehmenden eingeschätzt.

Zur Messung kristalliner Intelligenz wurde auf den umfangreichen Itempool des BEFKI-Projekts (Berliner Test zur Erfassung Fluider und Kristalliner Intelligenz; Wilhelm/Schroeders/Schipolowski, in Vorbereitung; Wilhelm/Schipolowski 2010) zurückgegriffen, mit dem deklaratives Wissen in 16 verschiedenen Domänen erfasst werden kann. Im Einzelnen wird naturwissenschaftliches (Physik, Chemie, Biologie, Medizin, Geografie, Technologie), geisteswissenschaftliches (Literatur, Kunst, Musik, Religion, Philosophie) und sozialwissenschaftliches Wissen erfragt (Geschichte, Recht, Politik, Wirtschaft, Finanzen). Die Auswahl der Wissensbereiche orientierte sich an der empirisch begründeten Klassifikation von Ackerman (2000; Rolfhus/Ackerman 1999). Für die aktuelle Studie wurden insgesamt 32 Wissensitems anhand inhaltlicher und psychometrischer Kriterien aus dem Gesamtitempool ausgewählt. Konkret wurden zwei Items aus jedem der 16 Wissensbereiche eingesetzt, wobei eines der beiden Items zu jedem Bereich von geringer bis mittlerer Schwierigkeit war (entwickelt für Personen ohne Schulabschluss, mit Hauptschul- oder Mittlerem Schulabschluss), das andere von hoher Schwierigkeit (entwickelt für Personen, die über die Hochschulreife verfügen bzw. diese anstreben). Die psychometrische Eignung der Items wurde anhand von Vorinformationen aus verschiedenen Erhebungen sichergestellt (Schipolowski/Schroeders/Wilhelm 2008; Schroeders/Schipolowski/Wilhelm 2010; Schroeders/Schipolowski/Nelles/Wilhelm 2011). Ausschlaggebend war dabei neben den Informationen zur Itemschwierigkeit insbesondere eine hohe positive Trennschärfe in den genannten Voruntersuchungen. Alle Wissensitems waren ausschließlich textbasiert und hatten ein Multiple-Choice-Format mit vier Antwortalternativen, von denen genau eine die richtige Lösung darstellte.

Zusätzlich zur kristallinen Intelligenz wurden weitere psychologische Konstrukte erfasst. Der BFI-10 (Rammstedt/John 2007) ermöglichte die Erfassung der Big-Five-Persönlichkeitsdimensionen Neurotizismus (N), Extraversion (E), Offenheit (O), Verträglichkeit (V) und Gewissenhaftigkeit (G) anhand von jeweils zwei (N, E, O, G) bzw. drei Items (V). Die Items bestanden aus Aussagen, deren Zutreffen von den Teilnehmenden auf einer fünfstufigen Ratingskala eingeschätzt wurde. Mit dem VOC-T (Ziegler/Kemper/Rammstedt 2013) wurde zudem ein Maß für die Selbsteinschätzung des eigenen Wissens eingesetzt. Zu insgesamt 12 verschiedenen Begriffen aus den Natur-, Geistes- und Sozialwissenschaften sowie dem handwerklichen Bereich gaben die teilnehmenden Personen anhand einer siebenstufigen Ratingskala an, wie vertraut sie mit dem jeweiligen Begriff oder Konzept sind. Neben diesen real existierenden Begriffen enthält der VOC-T zusätzlich drei fiktive Begriffe.

Weitere eingesetzte Skalen dienten der Erfassung von Lebenszufriedenheit, politischer Partizipation, Werten, Kontrollüberzeugungen, Selbstwirksamkeitser-

wartungen, Impulsivität und dem Gesundheitszustand. Da diese Skalen nur in das Imputations-, nicht jedoch in das Analysemodell einbezogen wurden, werden sie hier nicht näher beschrieben.

## 2.3 Durchführung

Die Erhebung der Daten erfolgte im Zeitraum 2. Mai bis 23. Juni 2011 durch ein beauftragtes Erhebungsinstitut. Geschulte Interviewer suchten die Studienteilnehmerinnen und -teilnehmer zu vorab vereinbarten Terminen in ihren Wohnungen auf, um die Befragung bzw. Testung durchzuführen. Die soziodemographischen Angaben und die Antworten auf die Persönlichkeitsitems wurden vom Interviewer erfragt und in eine Eingabemaske am Notebook eingegeben (CAPI, computer assisted personal interview). Den $g_c$-Test bearbeiteten die Teilnehmenden selbstständig am Notebook (CASI, computer assisted self-interview). In beiden Fällen war die Abfolge der Fragen durch ein Skript vorgegeben, um einen standardisierten Ablauf zu gewährleisten. Beim $g_c$-Test wurden immer vier Fragen gleichzeitig auf dem Bildschirm dargestellt; um zur nächsten Bildschirmseite zu gelangen, musste die teilnehmende Person zunächst alle vier Fragen der aktuellen Seite beantworten (ggf. durch Raten). Für die Bearbeitung der 32 $g_c$-Items war ein Zeitlimit von 10 Minuten vorgegeben. Bei Erreichen des Zeitlimits brach der Wissenstest automatisch ab und es wurde mit dem nächsten Fragenkomplex fortgefahren. Die Dauer des gesamten Interviews betrug im Durchschnitt 43 Minuten ($SD$ = 13).

## 2.4 Datenaufbereitung und Auswertungsverfahren

Zur Aufbereitung der gewichteten Stichprobendaten wurden im ersten Schritt anhand der vorliegenden demographischen Angaben verschiedene Indizes gebildet. Als Index zur formalen Bildung wurde die ISCED-97 (International Standard Classification of Education; UNESCO 1997) genutzt. Dabei wurden die Angaben zum höchsten erreichten Schulabschluss sowie zum Ausbildungsabschluss in einer ordinalskalierten Variable mit 6 Stufen zusammengeführt (vgl. Statistisches Bundesamt 2010: 79). Auf der niedrigsten Stufe (ISCED 1) befinden sich demnach Personen ohne Schul- und Ausbildungsabschluss, während die höchste Stufe (ISCED 6) mit Personen besetzt ist, die die allgemeine oder Fachhochschulreife besitzen und nach erfolgreichem Hoch- bzw. Fachhochschulabschluss zusätzlich einen weiterführenden akademischen Grad (Promotion) erlangt haben. Als Index des sozioökonomischen Status wurde der ISEI (International Socio-Economic Index of Occupational Status; Ganzeboom/De Graaf/Treiman 1992) verwendet. Der ISEI beruht auf Anga-

ben zur Berufstätigkeit und der Prämisse, dass diese Informationswert bezüglich Einkommen und Bildung besitzt, die wiederum „die Hauptquellen der Macht in modernen Gesellschaften" sind (Ganzeboom et al. 1992: 9). Konkret werden einzelnen Berufen Statuswerte zugeordnet, die auf internationalen Daten zum Einkommen und zur Bildung beruhen, über die Ausübende dieser Berufstätigkeiten typischerweise verfügen. Auf dieser Basis ergibt sich eine Skala von 16 (niedriger Status; etwa Reinigungskräfte) bis 90 (hoher Status; Richter). Ein Weg zur Ermittlung der ISEI-Werte ist die Kodierung von Berufstätigkeiten nach der ISCO (International Standard Classification of Occupations; ILO 2007). Im vorliegenden Fall wurde ein weniger aufwändiges Vorgehen gewählt, das auf Angaben zur beruflichen Stellung beruht, von der nach Wolf (1995) ebenfalls auf den ISEI geschlossen werden kann (vgl. Statistisches Bundesamt 2010). Nach diesem Vorgehen ergibt sich für die aktuelle Studie eine vereinfachte ISEI-Skala mit 13 verschiedenen Ausprägungen. Die Bildung des ISEI wurde für die Teilnehmenden selbst sowie für deren Eltern durchgeführt; in letzterem Fall wird für die folgenden Analysen der höchste der beiden elterlichen ISEI-Werte (HISEI) verwendet. Das Haushaltsnettoeinkommen wurde auch direkt erfragt, wobei zum Teil Freitextantworten, überwiegend jedoch Angaben in Form von Einkommenskategorien vorlagen. Freitextanworten wurden auf die vorliegenden 24 Kategorien rekodiert, um ein einheitliches Antwortformat zu erhalten.

Für die Skalen zu den psychologischen Konstrukten wurden Summen- oder Mittelwerte so gebildet, wie dies von den Autoren der jeweiligen Instrumente vorgeschlagen wurde. Somit lag für die folgenden Analysen jeweils ein Wert für jede der fünf Big-Five-Dimensionen vor. Als Indikator des selbstberichteten Wissens wurde ein Gesamtwert über alle 12 Items des VOC-T genutzt, die sich auf real existierende Begriffe beziehen; die fiktiven Begriffe wurden nicht einbezogen[2].

Die Items zur Messung der kristallinen Intelligenz wurden zunächst in richtig versus falsch beantwortet rekodiert. Anschließend wurde ein Summenwert über alle 32 Items als Schätzer der Personenfähigkeit auf der Gesamtskala berechnet. Auf Basis der 32 Wissensitems der Gesamtskala erfolgte im nächsten Schritt die Itemselektion für die Kurzskala nach folgenden Kriterien:

- Um einen flexiblen Einsatz der Kurzskala in der Umfrageforschung zu ermöglichen, sollte deren Bearbeitungszeit bei 5 Minuten liegen. Dies entspricht der geschätzten Bearbeitungszeit von 12 Items.

2 Die fiktiven Begriffe des VOC-T, engl. *foils*, können zur Berechnung weiterer Kennwerte herangezogen werden, die jedoch im vorliegenden Beitrag nicht berücksichtigt werden.

- Zur bestmöglichen Erhaltung der inhaltlichen Breite der Wissensmessung sollte einerseits die Dreiteilung in natur-, geistes- und sozialwissenschaftliches Wissen beibehalten werden, andererseits sollten möglichst viele der 16 Wissensbereiche der Gesamtskala auch in der Kurzskala enthalten sein.
- Um Boden- und Deckeneffekte zu minimieren, sollten die ausgewählten Items einen großen Schwierigkeitsbereich abdecken. Die relative Lösungshäufigkeit sollte jedoch stets oberhalb der Ratewahrscheinlichkeit von .25 liegen.
- Es wurde ein einfaktorielles Messmodell mit guter Modellpassung und Itemladungen (Trennschärfen) von .50 oder höher angestrebt. In keinem Fall sollten Ladungen < .30 auftreten.
- Die Kurzskala sollte ähnliche Beziehungen zu Personen- und Umweltmerkmalen sowie anderen psychologischen Konstrukten aufweisen wie die Gesamtskala.

Nach erfolgter Itemselektion wurde für die Kurzskala mit 12 Items ebenfalls ein Summenwert berechnet.

Um Einschränkungen bei der Teststärke sowie Verzerrungen durch nicht zufällig fehlende Informationen[3] zu minimieren, wurden fehlende Datenpunkte im Wissenstest sowie in allen Kovariaten imputiert (Lüdtke/Robitzsch/Trautwein/Köller 2007). Der Anteil fehlender Werte bei den 32 Wissensitems betrug im Mittel pro Item 5.8% (*SD* = 7.0%, Spannweite 0.4% bis 20.8%); die vorliegenden Fallzahlen für die Kovariaten gehen aus der Ergebnistabelle hervor (vgl. Tabelle 4, Spalte $N_{vi}$). Speziell bei den Wissensitems, für die Datenpunkte fast ausschließlich aufgrund der Zeitbegrenzung fehlten, ermöglichte die Imputation eine Minimierung konstruktirrelevanter Varianz (etwa interindividuelle Unterschiede in mentaler Geschwindigkeit; Danthiir/Roberts/Schulze/Wilhelm 2004). Konkret wurde das Verfahren der multiplen Imputation mit 100 Replikationen genutzt (Graham/Olchowski/Gilreath 2007). Um die fehlenden Werte möglichst zuverlässig schätzen zu können, wurden alle verfügbaren Kovariaten, also Personen- und Umweltmerkmale sowie psychologische Konstrukte, in das Imputationsmodell einbezogen (Collins/Schafer/Kam 2001). Die im Ergebnisteil berichteten Statistiken und Koeffizienten sind Mittelwerte über alle Replikationen. Bei der Ermittlung von Standardfehlern wurde die Streuung zwischen den Replikationen berücksichtigt.

Zur Berechnung von Kovarianzen bzw. Korrelationen kamen Verfahren zum Einsatz, die dem Skalenniveau der einbezogenen Variablen Rechnung tragen.

3 Als „nicht zufällig fehlende Informationen" sind solche Ausfälle zu betrachten, die von der Ausprägung der fraglichen Variable selbst und/oder anderen beobachteten Variablen abhängig sind; vgl. die Definition von *Missing Not At Random* (MNAR) und *Missing At Random* (MAR) bei Rubin (1976).

Tabelle 1 Schwierigkeiten und Trennschärfen der Kurzskala-Items

| Item-Nr. | Itembezeichner | Wissensbereich | Wissensdomäne | *p* | λ |
|---|---|---|---|---|---|
| 1 | med | Medizin | Naturw. | .88 | .58 |
| 2 | rel | Religion | Geistesw. | .69 | .67 |
| 3 | geo | Geografie | Naturw. | .78 | .58 |
| 4 | kun | Kunst | Geistesw. | .44 | .64 |
| 5 | bio | Biologie | Naturw. | .33 | .45 |
| 6 | pol | Politik | Sozialw. | .62 | .52 |
| 7 | phi | Philosophie | Geistesw. | .40 | .51 |
| 8 | phy | Physik | Naturw. | .44 | .35 |
| 9 | lit | Literatur | Geistesw. | .51 | .56 |
| 10 | fin | Finanzen | Sozialw. | .77 | .58 |
| 11 | rec | Recht | Sozialw. | .64 | .39 |
| 12 | ges | Geschichte | Sozialw. | .55 | .46 |

*Anmerkungen: Naturw.: Naturwissenschaften; Geistesw.: Geisteswissenschaften; Sozialw.: Sozialwissenschaften; p: relative Lösungshäufigkeit; λ: Itemladung im einfaktoriellen Messmodell (vgl. Abbildung 1, Modell b).*

Entsprechend der kategorialen (dichotomen) Natur der rekodierten Wissensitems wurden Messmodelle mittels konfirmatorischer Faktorenanalyse unter Verwendung des WLSMV-Schätzers in Mplus 6.1 (Muthén/Muthén 1998-2010) berechnet. Zur Beurteilung der Modellpassung wurden neben dem Chi-Quadrat-Wert und den Freiheitsgraden weitere gebräuchliche Fitindizes wie CFI (Comparative fit index), RMSEA (Root mean square error of approximation) und WRMR (Weighted root mean square residual) herangezogen. Nach Yu (2002) zeichnen sich Modelle mit kategorialen Daten und guter Passung durch folgende Werte aus: CFI ≥ .96, RMSEA ≤ .05 und WRMR ≤ .95. Alle Analysen erfolgten unter Verwendung der Fallgewichte.

# 3 Ergebnisse

## 3.1 Itemstatistiken

Die Entwicklung der Kurzskala erfolgte auf Basis des vorliegenden Datensatzes durch Itemselektion aus der 32 Items umfassenden Gesamtskala nach den oben genannten Kriterien (vgl. Abschnitt 2.4). In Tabelle 1 sind die Schwierigkeiten, die Itemladungen (äquivalent zu Trennschärfen in der klassischen Testtheorie) und die

inhaltliche Zuordnung zu den drei breiten Wissensdomänen bzw. den einzelnen Wissensbereichen für die ausgewählten Items angegeben. Werden die entsprechenden Summenscores herangezogen, korreliert die Kurzskala mit der Gesamtskala zu $r = .91$.

## 3.2 Messmodelle und Reliabilität

Im Folgenden werden zwei konkurrierende Messmodelle für die Kurzskala gegenübergestellt und hinsichtlich ihrer Passung verglichen. Im dreifaktoriellen Modell werden drei Faktoren gemäß der drei breiten Wissensdomänen Natur-, Geistes- und Sozialwissenschaften spezifiziert. Auf jedem der drei Faktoren laden dabei nur die Items aus der entsprechenden Domäne. Da kristalline Intelligenz als übergeordnetes Fähigkeitskonstrukt Wissen aller Bereiche umfasst, ist von positiven Korrelationen zwischen den drei domänenspezifischen Faktoren auszugehen. Das einfaktorielle Messmodell repräsentiert kristalline Intelligenz hingegen mit einem einzigen Faktor, auf dem alle 12 Items der Kurzskala laden. Eine weitere Untergliederung in verschiedene Wissensdomänen wird nicht modelliert. Das einfaktorielle Modell stellt somit einen Spezialfall des komplexeren dreifaktoriellen Modells dar, weshalb zum Modellvergleich auch die Differenz der Chi-Quadrat-Werte formal auf Signifikanz getestet werden kann (Schulze 2004). Die beiden Messmodelle sind in Abbildung 1 dargestellt; die Passung der Modelle kann Tabelle 2 entnommen werden.

Das dreifaktorielle Modell (a) weist eine signifikant bessere Passung auf ($\Delta\chi^2(N = 1.134, 3) = 15.2$, $p = .002$), was sich auch an den anderen in Tabelle 2 ausgewiesenen Fitindizes zeigt. Der Unterschied ist allerdings gering; auch für das einfaktorielle Modell (b) ergibt sich eine zufriedenstellende Passung. Übereinstimmend mit dem geringen Unterschied in der Modellpassung zeigen sich im dreifaktoriellen Modell hohe messfehlerbereinigte Korrelationen zwischen den drei Wissensdomänen, die zwischen .80 und .96 liegen.

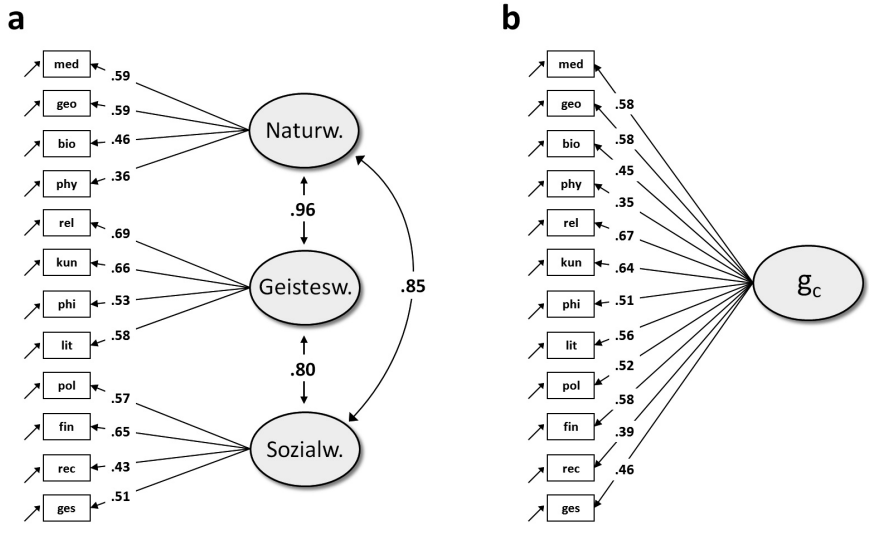
Für die folgenden Analysen wird das einfaktorielle Messmodell verwendet, da es eine zufriedenstellende Passung aufweist, die nur geringfügig schlechter ist als die Passung des dreifaktoriellen Modells. Insbesondere ist der inhaltliche Mehrwert des dreifaktoriellen Modells fraglich, da die drei breiten Wissensdomänen hoch zusammenhängen und in der Kurzskala nur durch jeweils vier Items repräsentiert werden, wodurch eine substanzielle Interpretation dieser spezifischeren Faktoren erschwert wird. Für das einfaktorielle Messmodell ergibt sich eine zufriedenstellende Reliabilität der latenten Variable von $\omega = .82$ (McDonald 1999) bzw. $\alpha = .81$ (Zumbo/Gadermann/Zeisser 2007). Die Reliabilität des manifesten Summenscores

Tabelle 2 Fitstatistiken konkurrierender Messmodelle

| Modell | $\chi^2$ | *df* | CFI | RMSEA | WRMR |
|---|---|---|---|---|---|
| 3 Faktoren | 94.9 | 51 | .973 | .027 | 0.97 |
| 1 Faktor | 110.8 | 54 | .965 | .030 | 1.06 |

*Anmerkungen: $\chi^2$: Chi-Quadrat-Wert; df: Anzahl der Freiheitsgrade; CFI: Comparative fit index; RMSEA: Root mean square error of approximation; WRMR: Weighted root mean square residual.*

Abbildung 1 Dreifaktorielles (a) und einfaktorielles Messmodell (b) der Kurzskala




*Anmerkungen: Naturw.: Naturwissenschaften; Geistesw.: Geisteswissenschaften; Sozialw.: Sozialwissenschaften, $g_c$: kristalline Intelligenz; zu den Itembezeichnern siehe Tabelle 1.*

liegt bei .70 (Raykov/Dimitrov/Asparouhov 2010). Die entsprechenden Werte für die Gesamtskala mit 32 Items liegen mit .91 ($\omega/\alpha$) bzw. .84 (Skalenreliabilität nach Raykov et al. 2010) wegen der höheren Itemzahl erwartungsgemäß höher.

## 3.3 Überprüfung von Boden- und Deckeneffekten

Da bei Kurzskalen nur wenige Items eingesetzt werden, besteht im Vergleich zu umfangreicheren Skalen ein höheres Risiko von Boden- bzw. Deckeneffekten, d.h. mangelnder Differenzierungsfähigkeit innerhalb von Personengruppen mit sehr niedriger bzw. sehr hoher Fähigkeitsausprägung. Um zu überprüfen, ob solche

Effekte vorliegen, wird im Folgenden zunächst die Verteilung des Summenwerts der Kurzskala in der Gesamtpopulation betrachtet (vgl. Abbildung 2). Zudem erfolgt eine Betrachtung zweier Subpopulationen mit geringer bzw. hoher Schulbildung (vgl. Abbildung 3). Die Subpopulation mit „geringer Schulbildung" umfasst sowohl Personen ohne Schulabschluss als auch Personen mit Hauptschulabschluss (bzw. Äquivalent), die in der Regel nach der achten oder neunten Klasse die allgemeinbildende Schule verlassen haben. Ihr Anteil an der erwachsenen deutschen Wohnbevölkerung beträgt gemäß der vorliegenden Erhebung 45%. Personen mit „hoher Schulbildung" im Sinne der hier vorgenommenen Analyse verfügen demgegenüber über eine fachgebundene oder allgemeine Hochschulreife bzw. Fachhochschulreife (26% der Gesamtpopulation), die typischerweise nach 12 oder 13 Jahren Schulbesuch erworben wurde.

Tabelle 3 gibt die Kennwerte der Verteilungen wieder. Für die Gesamtpopulation zeigen sich keine nennenswerten Boden- oder Deckeneffekte. Zwar liegt eine gering negative Schiefe vor, d.h. am oberen Ende der Fähigkeitsskala ist die Differenzierungsfähigkeit minimal geringer. Dies ist jedoch praktisch kaum bedeutsam: Der Anteil der Personen, die alle Wissensitems der Kurzskala richtig beantworten können, liegt in der Gesamtpopulation unter vier Prozent; im Mittel werden etwa sieben Items richtig gelöst. Ein differenzierteres Bild ergibt sich erwartungsgemäß für die zwei betrachteten Subpopulationen, die sich in der mittleren Lösungshäufigkeit deutlich unterscheiden. Während Personen mit „geringer Schulbildung" im Mittel weniger als sechs Items richtig beantworten, liegt der Mittelwert in der Subpopulation mit „hoher Schulbildung" etwas unter neun Items. Konkret beträgt die standardisierte Mittelwertdifferenz zwischen den beiden Subpopulationen $d = 1.12$ Standardabweichungseinheiten. In der relativ großen Gruppe der Personen ohne Schulabschluss bzw. mit Hauptschulabschluss liegt kein Bodeneffekt vor, hier liegt die Schiefe nahe null. Für die kleinere Gruppe der Personen mit Hochschulreife tritt erwartungsgemäß ein Deckeneffekt auf. Selbst in dieser sehr leistungsstarken Personengruppe liegt jedoch der Anteil derer, die alle Items richtig beantworten, unter 10 Prozent.

## 3.4 Zusammenhänge mit Kriterien

Ein wesentlicher Aspekt bei der Entwicklung und Beurteilung von Kurzskalen besteht in deren Beziehungen zu relevanten Personen- und Umweltmerkmalen sowie anderen psychologischen Konstrukten. Diese sollen einerseits im Einklang mit der Literatur stehen (Konstruktvalidierung). Andererseits sollen die für die Kurzskala ermittelten Beziehungen möglichst den Befunden für die ungekürzte

Abbildung 2 Verteilung des Summenwerts der Kurzskala in der Gesamtpopulation

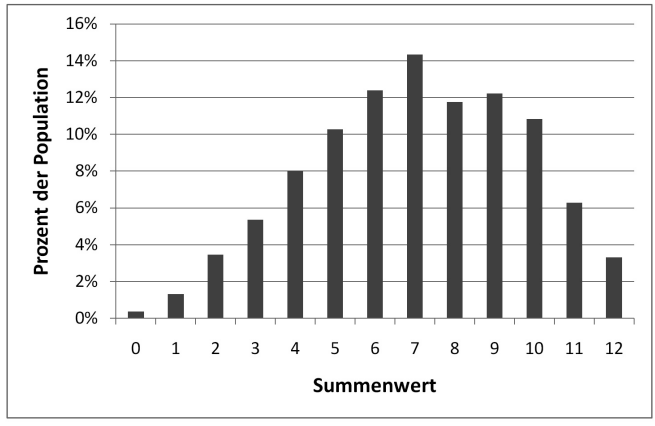



Abbildung 3 Verteilung des Summenwerts der Kurzskala in zwei Subpopulationen

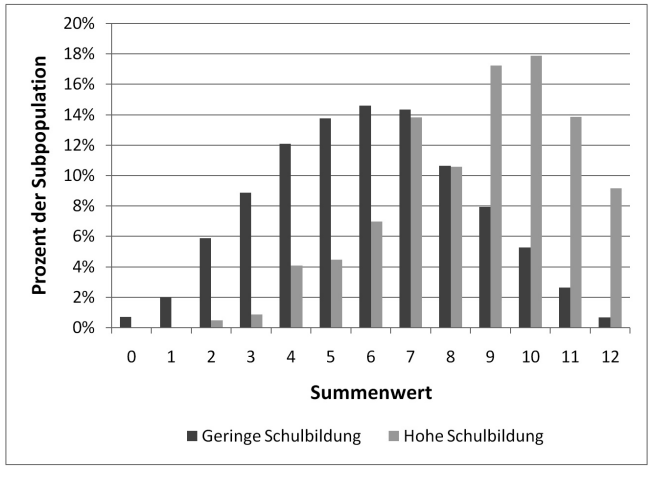



Tabelle 3 Kennwerte verschiedener Personenverteilungen für den Summenwert der Kurzskala

| Population | *N* | *M* | *SD* | Schiefe | Exzess |
|---|---|---|---|---|---|
| Gesamtpopulation | 1.134 | 7.04 | 2.66 | -0.20 | -0.64 |
| „Geringe Schulbildung" | 514 | 5.95 | 2.49 | +0.06 | -0.17 |
| „Hohe Schulbildung" | 290 | 8.66 | 2.27 | -0.52 | -0.38 |

*Anmerkungen: N: Stichprobengröße; M: arithmetisches Mittel; SD: Standardabweichung.*

Gesamtskala entsprechen, um sicherzustellen, dass die Itemselektion keine substanzielle Minderung der Konstruktvalidität zur Folge hat (Widaman/Little/Preacher/Sawalani 2011). Im Folgenden werden daher die Korrelationen der Kurzskala mit verschiedenen Kovariaten näher betrachtet und den entsprechenden Korrelationen der Gesamtskala mit 32 Wissensitems gegenübergestellt (vgl. Tabelle 4). Zu den betrachteten Personenvariablen zählen Geschlecht, Alter, formale Bildung und sozioökonomischer Status der Teilnehmenden. Darüber hinaus werden an dieser Stelle auch Merkmale des Haushalts der Befragten (Haushaltsnettoeinkommen) sowie des elterlichen Haushalts analysiert (sozioökonomischer Status der Eltern, Anzahl der Bücher im elterlichen Haushalt zur Jugendzeit der Teilnehmenden), die als Indikatoren für Umfang und Reichhaltigkeit der früheren oder aktuellen Lernumwelt angesehen werden. Mit Blick auf psychologische Konstrukte werden die Beziehungen zu den fünf Hauptdimensionen der Persönlichkeit (Big Five) und zu selbstberichtetem Wissen untersucht. Bei der Interpretation der im folgenden dargestellten Beziehungen ist zu beachten, dass es sich um messfehlerbehaftete Korrelationen zwischen manifesten Variablen handelt. Zur Einordnung der Größe der Effekte kann eine Orientierung an Cohen (1988) erfolgen, der für Produkt-Moment-Korrelationen Werte um .10 als kleine Effekte, Werte um .30 als mittlere Effekte und Werte um .50 als große Effekte betrachtet. Wesentlicher ist jedoch der Vergleich der hier ermittelten Zusammenhänge mit den in der Einleitung dargestellten Erwartungen.

In Übereinstimmung mit der Literatur zeigen Männer etwas höhere Wissensleistungen als Frauen. Die standardisierte Mittelwertdifferenz beträgt $d$ = .30 und ist somit inhaltlich bedeutsam, obgleich niedriger als von Ackerman et al. (2001) berichtet[4]. Wie erwartet wird in der hier untersuchten Erwachsenenpopulation kein bedeutsamer Alterseffekt beobachtet. Deklaratives Wissen weist eine hohe positive Korrelation mit dem ISCED-97 als Indikator formaler Bildung auf, der sowohl Schul- als auch Ausbildungsabschlüsse berücksichtigt: Für kein anderes hier untersuchtes Personen- oder Umweltmerkmal wurden höhere Korrelationen gefunden. Eine ebenfalls starke Beziehung zeigt sich zu dem auf der ISEI-Skala quantifizierten sozioökonomischen Status der Teilnehmenden. Auch für die anderen in Tabelle 4 genannten Personen- und Haushalts-

4 Für Männer und Frauen kann von Messinvarianz ausgegangen werden, d.h. die Kurzskala misst in beiden Gruppen dasselbe Konstrukt mit vergleichbarer Genauigkeit. Ein entsprechend restringiertes Multigruppenmodell unter Annahme strikter Messinvarianz zeigte sowohl für die Kurzskala mit 12 Items als auch für die Gesamtskala mit 32 Wissensitems eine befriedigende Passung (Werte für die Kurzskala: $\chi^2$ = 203.7, $df$ = 130, RMSEA = .032, CFI = .953, WRMR = 1.548). Der latente Mittelwertunterschied betrug 0.32 *SD* zugunsten der Männer.

Tabelle 4 Korrelationen der Kurz- und Gesamtskala mit verschiedenen Personen- und Haushaltsmerkmalen sowie psychologischen Konstrukten

| | | Kurzskala | | Gesamtskala | |
|---|---|---|---|---|---|
| Variable | $N_{vi}$ | *r* | *SE* | *r* | *SE* |
| Geschlecht[1] | 1.134 | -.15 | .03 | -.15 | .03 |
| Alter | 1.134 | .01[n.s.] | .03 | .00[n.s.] | .03 |
| ISCED-97 | 1.091 | .49 | .03 | .51 | .03 |
| ISEI | 388 | .44 | .04 | .45 | .04 |
| HISEI Eltern | 1.082 | .25 | .03 | .25 | .03 |
| Einkommen | 638 | .29 | .04 | .30 | .04 |
| Anzahl Bücher | 1.101 | .30 | .03 | .33 | .03 |
| selbstber. Wissen[2] | 1.134 | .52 | .03 | .55 | .03 |
| Neurotizismus[3] | 1.104 | -.10 | .03 | -.15 | .03 |
| Extraversion[3] | 1.104 | .07[a] | .04 | .12 | .04 |
| Offenheit[3] | 1.104 | .21 | .03 | .25 | .03 |
| Gewissenhaftigkeit[3] | 1.104 | .07[a] | .03 | .09 | .03 |
| Verträglichkeit[3] | 1.104 | -.02[n.s.] | .03 | -.02[n.s.] | .03 |

*Anmerkungen: N = 1.134. [1] 0 = männlich, 1 = weiblich; [2] selbstberichtetes Wissen (VOC-T Treffer); [3] Big Five-Dimensionen; [a] p < .05; [n.s.] nicht signifikant; $N_{vi}$: Fallzahl vor der Imputation; r: punkt-biseriale Korrelation (Geschlecht), polyseriale Korrelation (ISCED-97, Bücher), Produkt-Moment-Korrelation (alle anderen Variablen); SE: Standardfehler; ISCED-97: International Standard Classification of Education, Fassung 1997; ISEI: International Socio-Economic Index of Occupational Status; HISEI Eltern: Höchster ISEI-Wert der beiden Elternteile des Teilnehmenden. Sofern nicht anders gekennzeichnet, sind alle Korrelationen signifikant von null verschieden (p < .01).*

merkmale liegen inhaltlich bedeutsame positive Korrelationen mit Wissen vor, die jedoch niedriger als die Zusammenhänge mit ISCED-97 und ISEI ausfallen.

Mit Blick auf die psychologischen Konstrukte liegt erwartungsgemäß eine vergleichsweise hohe Korrelation mit selbstberichtetem Wissen vor. Für die fünf Persönlichkeitsdimensionen ergibt sich ein differenziertes Ergebnismuster, das im Wesentlichen mit den in der Literatur berichteten Befunden übereinstimmt. Die Offenheitsdimension weist im Vergleich mit den anderen Big Five-Dimensionen die vom Betrag her höchste Korrelation mit Wissen auf. Auch die gering positive Korrelation mit Extraversion, die gering negative Beziehung zu Neurotizismus sowie die nicht signifikante Korrelation mit Verträglichkeit entsprechen den metaanalytisch gewonnenen Ergebnissen von Ackerman und Heggestad (1997). Eine Abweichung lässt sich allein für Gewissenhaftigkeit feststellen: Während Ackerman und Heggestad (1997) hier eine negative Korrelation mit Wissen bzw. eine nicht signifikante

Korrelation mit kristalliner Intelligenz berichten, wird in der aktuellen Analyse eine gering positive Beziehung zwischen Wissen und Gewissenhaftigkeit gefunden. Für alle berichteten Kovariaten ergeben sich im Vergleich von Kurz- und Gesamtskala sehr ähnliche Korrelationen.

# 4 Diskussion

Da der Umfrageforschung im deutschen Sprachraum bis dato nur wenige frei verfügbare und erprobte Instrumente zur Erfassung kognitiver Fähigkeiten vorliegen, war das Ziel dieses Beitrags die Entwicklung einer Kurzskala zur Messung kristalliner Intelligenz ($g_c$) und die Prüfung ihrer psychometrischen Eigenschaften. Die hier beschriebene Kurzskala BEFKI GC-K umfasst 12 Items, die in Übereinstimmung mit der Definition von $g_c$ durch Cattell (1971) bzw. Carroll (1993) deklaratives Wissen aus ebensovielen Bereichen der Natur-, Geistes- und Sozialwissenschaften erfassen. Somit wird innerhalb von fünf Minuten Bearbeitungszeit ein möglichst breites Wissensspektrum berücksichtigt. Durch Einbezug von Items unterschiedlicher Schwierigkeit gelang es, trotz der vergleichsweise geringen Itemanzahl Boden- und Deckeneffekte in der erwachsenen deutschen Wohnbevölkerung fast vollständig zu vermeiden. Lediglich in der besonders leistungsstarken Subpopulation mit Hochschulabschluss bzw. Fachhochschulabschluss trat ein geringer Deckeneffekt auf. Weiterhin weist die Kurzskala eine gute Reliabilität auf, welche gängige Standards für die Forschung erfüllt. Die zeitliche Stabilität der Testwerte konnte aufgrund der querschnittlichen Natur der Erhebung nicht geprüft werden. Daten einer mit Schülerinnen und Schülern der Mittelstufe durchgeführten Längsschnittuntersuchung mit Wissensitems aus dem BEFKI-Pool legen jedoch eine befriedigende zeitliche Stabilität nahe (Wilhelm et al., in Vorbereitung).

Sowohl für ein einfaktorielles Messmodell als auch für ein mehrdimensionales Modell, das drei korrelierte Faktoren gemäß der Unterscheidung zwischen Natur-, Geistes- und Sozialwissenschaften spezifiziert, ergab sich eine zufriedenstellende Passung bei statistisch signifikanter Überlegenheit des dreifaktoriellen Modells. Die Bevorzugung des eindimensionalen Modells im vorliegenden Beitrag beruht auf mehreren Argumenten: Erstens ist zu berücksichtigen, dass der $\chi^2$-Differenztest von der Stichprobengröße abhängig ist und daher in der vorliegenden, vergleichsweise großen Stichprobe auch inhaltlich unbedeutende Unterschiede als signifikant ausgewiesen werden. Zweitens ist das einfaktorielle Modell insofern theoretisch fundiert, als Wissen aus verschiedenen Bereichen in konsensualen Intelligenzstrukturtheorien einem gemeinsamen, bereichsübergreifenden $g_c$-

Faktor untergeordnet ist (Horn/Noll 1997; Carroll 1993, 2003; McGrew 2009). Dies wird auch empirisch durch die sehr hohen Korrelationen zwischen den Wissensfaktoren im dreidimensionalen Modell gestützt. Drittens ist zu berücksichtigen, dass die spezifischeren Faktoren des dreidimensionalen Modells mit nur jeweils vier Items extrem schmal operationalisiert sind. Eine inhaltliche Interpretation dieser Faktoren wäre fragwürdig, da die Itemstichprobe aufgrund dieser sehr geringen Itemzahl keine Repräsentativität für die jeweilige Wissensdomäne beanspruchen kann (Ackerman 2000). Ferner ist die im Vergleich zu einem Gesamtwert geringere Reliabilität von Subskalen zu berücksichtigen (Sinharay 2010). Im Extremfall kann dies dazu führen, dass der Gesamttestwert einen besseren Prädiktor der Personenfähigkeit auf einer Subdimension darstellt als der entsprechende Subskalenwert (Sinharay/Haberman/Puhan 2007). Selbst bei vermeintlich ausreichender Reliabilität einer Subskala kann diese durch die übergeordnete Fähigkeit bedingt sein statt durch die spezifische Subdimension (Reise/Moore/Haviland 2010).

Die Validität der $g_c$-Kurzskala wurde anhand der Korrelationen des Gesamtwertes mit verschiedenen Personen- und Umweltmerkmalen einerseits sowie ausgewählten psychologischen Konstrukten andererseits überprüft, wobei diese Beziehungen fast durchgängig mit den auf Basis der Fachliteratur formulierten Erwartungen übereinstimmten. Die Höhe der untersuchten Effekte bzw. Zusammenhänge lag für die Kurzskala mit 12 Items nur geringfügig unter den entsprechenden Werten für die ungekürzte Wissensskala mit 32 Items; die Abweichungen lassen sich durch die etwas geringere Reliabilität der Kurzskala erklären. Im Einzelnen wurde übereinstimmend mit den oben zitierten Befunden ein Geschlechtsunterschied im deklarativen Wissen zugunsten von Männern gefunden. Dies stellt keine Verzerrung und somit unfaire Messung dar, sondern repliziert den Befund von Ackerman et al. (2001), dass bei einem umfassenden Sampling von Items aus vielen verschiedenen Wissensbereichen in den meisten dieser Bereiche ein Wissensvorsprung der Männer zu beobachten ist. Dass der Mittelwertunterschied in der hier vorgestellten Skala weniger deutlich ausfällt, könnte auf kulturelle Unterschiede zurückzuführen sein oder darin begründet liegen, dass andere Studien teilweise deutlich größere Itemmengen eingesetzt haben, einschließlich sehr spezifischer Wissensfragen, die als berufsspezifische Expertise einzuordnen sind (vgl. den $g_{kn}$-Faktor im CHC-Modell; McGrew 2009). In der vorliegenden Kurzskala wurde hingegen auf derartige Items zu hochspezifischem Wissen verzichtet, da sie nur innerhalb bestimmter Subpopulationen funktionieren. Dies ist auch ein Grund dafür, dass kein substanzieller Zusammenhang zwischen der $g_c$-Kurzskala mit dem Alter festgestellt werden konnte. Mit Blick auf die hohe Stabilität kristalliner Intelligenz im Erwachsenenalter war ein kleiner oder nicht signifikanter Effekt erwartet

worden. Zwar werden in der Literatur Wissenszuwächse auch im jungen und mittleren Erwachsenenalter berichtet (Ackerman 2000), diese beziehen sich aber auf Expertise im Sinne von Wissen, das erst im Erwachsenenalter erworben werden kann, beispielsweise im Rahmen der Berufsausübung (Ackerman 1996). Zum anderen umfasst die untersuchte Population neben jungen Erwachsenen auch Erwachsene in sehr hohem Alter, für die in der Literatur biologisch bedingt abnehmende Leistungen auch in Tests kristalliner Fähigkeiten berichtet werden (Li 2003).

Erwartungskonform sind ebenfalls die substanziellen positiven Beziehungen der Kurzskala mit Indikatoren der formalen Bildung und der Reichhaltigkeit der Lernumwelt, die im sozioökonomischen Status zum Ausdruck kommt. Sowohl ISCED-97 als auch ISEI der Teilnehmenden hängen mit dem über einen langen Zeitraum in einer Vielzahl von Bereichen erworbenen Wissen zusammen. Die vergleichsweise niedrigeren Korrelationen mit dem sozioökonomischen Status der Eltern, dem Haushaltsnettoeinkommen sowie der Anzahl der Bücher im elterlichen Haushalt zur Jugendzeit der Teilnehmenden sind insofern plausibel, als es sich dabei um eher indirekte bzw. inhaltlich weniger breite Indikatoren handelt, die begrenzte Aussagekraft bzgl. der Lernumwelt haben.

Mit Blick auf andere psychologische Konstrukte konnte gezeigt werden, dass eine relativ hohe positive Korrelation zwischen der Kurzskala und selbstberichtetem Wissen vorliegt. Dieser Befund überrascht nicht, da beide Skalen auf deklaratives Wissen abzielen und die Auswahl der erfassten Wissensbereiche in beiden Messverfahren auf der Systematik von Ackerman (2000) beruht. Dass kein perfekter Zusammenhang vorliegt, dürfte – neben der Tatsache, dass die berichteten Korrelationen messfehlerbehaftet sind – insbesondere an der unterschiedlichen Natur der Messverfahren liegen (Selbstauskünfte versus tatsächliches Wissen). Die Beziehungen der $g_c$-Kurzskala zu den Big Five zeigten ein Korrelationsmuster, das mit den metaanalytischen Befunden von Ackerman und Heggestad (1997) weitgehend übereinstimmt. Die höchste Korrelation wies die $g_c$-Kurzskala wie erwartet mit der Offenheitsdimension auf. Im Widerspruch zu den Ergebnissen von Ackerman und Heggestad (1997) steht lediglich die leicht positive Korrelation mit der Gewissenhaftigkeitsdimension. Hierbei ist jedoch zu bedenken, dass die diesbezügliche Datenbasis in der genannten Metaanalyse sehr klein war und ein gering positiver Zusammenhang zwischen Gewissenhaftigkeit und Wissensleistungen aus theoretischer Sicht erklärbar ist. Man denke etwa an die umfangreiche Literatur zur prädiktiven Validität von Gewissenhaftigkeit für den Berufserfolg (Barrick/Mount/Judge 2001). Befunde zu den Beziehungen der Kurzskala mit anderen $g_c$-Indikatoren sowie mit $g_f$ liegen bisher nur für Langformen aus Pilotierungs- und Normierungsstudien vor, die mit Schülerpopulationen durchgeführt wurden. Auf

der Ebene latenter Variablen konnte bei Schülerinnen und Schülern der achten bis zehnten Jahrgangsstufe eine hohe Korrelation zwischen deklarativem Wissen und Wortschatz ($\rho$ = .93) und eine ebenfalls substanzielle Beziehung zwischen Wissen und schlussfolgerndem Denken ($\rho$ = .80) gezeigt werden (Wilhelm et al., in Vorbereitung).

Bei der Interpretation der im vorliegenden Beitrag berichteten Befunde ist einschränkend zu berücksichtigen, dass die Itemselektion für die Kurzskala und deren Evaluation auf derselben Erhebung basieren. Zudem lag für den Wissenstest eine relativ strenge Zeitbegrenzung vor, die zu fehlenden Werten führte, da nicht alle teilnehmenden Personen den Test in der vorgegebenen Zeit abschließen konnten. Diese aufgrund der Zeitbegrenzung fehlenden Werte als Falschantworten zu werten, hätte eine Verzerrung des Skalenwerts durch konstruktirrelevante Varianz – etwa Unterschiede in mentaler Geschwindigkeit – zur Folge gehabt. Um diese Verzerrung zu vermeiden, wurden daher fehlende Werte unter Berücksichtigung aller verfügbaren Informationen imputiert; dadurch ist die größtmögliche Vergleichbarkeit mit einer Durchführung ohne Zeitmangel gewährleistet. Eine Bearbeitungzeit von fünf Minuten für die 12 Items der Kurzskala ist jedoch ausreichend, so dass eine Imputation fehlender Werte unter diesen Durchführungsbedingungen nicht erforderlich ist (stattdessen sollten ausgelassene Items wie Falschantworten ausgewertet werden). In zukünftigen Studien sollte geprüft werden, inwiefern die hier berichteten Befunde auf andere Stichproben und die veränderten Durchführungsbedingungen übertragbar sind.

## 5 Abschließende Bemerkungen

Die hier vorgestellte Kurzskala BEFKI GC-K zur Erfassung kristalliner Intelligenz stellt unseres Wissens das erste frei verfügbare Verfahren zur $g_c$-Messung dar, das sich aufgrund seiner Kürze und psychometrischen Effizienz für den Einsatz in der Umfrageforschung eignet. Die Skala ist sowohl in technologiebasierten Testungen (computer assisted personal interview/self-interview/web-interview) als auch bei herkömmlichen Papier-Stift-Testungen (Selbstausfüller) leicht anzuwenden; neben Einzeltestungen sind auch Gruppentestungen problemlos möglich. Einschränkend gilt hier, dass die Bewährung der Kurzskala mit anderen Testmedien bislang nicht geprüft wurde. Mit Blick auf die umfangreiche Literatur zu Testmedienvergleichen (vgl. etwa Mead/Drasgow 1993; Schroeders/Wilhelm 2010) sind für die $g_c$-Skala bei ansonsten gleichen Durchführungsbedingungen jedoch keine nennenswerten Testmedieneffekte zu erwarten. Gegenüber gängigen Bildungsindikatoren wie

der ISCED-97 hat die Wissensskala den Vorteil, dass tatsächliches Wissen direkt erhoben wird, anstatt es aus Abschlüssen zu erschließen. Letzteres ist insbesondere deshalb problematisch, da Schul- und Ausbildungsabschlüsse einerseits zwischen verschiedenen Bundesländern oder gar Staaten schwer vergleichbar sind (Neumann/Nagy/Trautwein/Lüdtke 2009) und formale Abschlüsse andererseits dem zeitlichen Wandel unterliegen. Zudem sind Selbstauskünfte wie Angaben zu Bildungsabschlüssen oder zur Berufstätigkeit im Gegensatz zu objektiven Leistungstests leicht verfälschbar (Ziegler/MacCann/Roberts 2011). Ungeachtet der Vorzüge der hier vorgestellten $g_c$-Skala ist jedoch zu berücksichtigen, dass es sich um eine Kurzskala handelt, die zur Wissensmessung in der sehr bildungsheterogenen erwachsenen Bevölkerung konstruiert wurde. Daher ist die inhaltliche Breite der Messung stark eingeschränkt und kann eine differenzierte Wissensdiagnostik nicht ersetzen. Stattdessen fokussiert die Kurzskala auf „Wissen, von dem angenommen werden kann, dass es von einem großen Teil der Population geteilt wird" (Ackerman 2003: 16). Mithilfe komplexer Testdesigns und Modellierungsmethoden ist es jedoch möglich, die Vorzüge von Kurzskalen mit den Vorteilen umfangreicher Messinstrumente zu kombinieren (Rhemtulla/Little 2012).

Die Kurzskala kann für nicht-kommerzielle Forschungszwecke kostenfrei eingesetzt werden. Eine Druckvorlage ist über folgenden Link verfügbar: http://befki.de/materialien (Kennwort: trgw34785bns)

## Literatur

Ackerman, P. L., 1996: A theory of adult intellectual development: Process, personality, interests, and knowledge. Intelligence 22: 227–257.

Ackerman, P. L., 2000: Domain-Specific Knowledge as the „Dark Matter" of Adult Intelligence: Gf/Gc, Personality and Interest Correlates. Journal of Gerontology: Psychological Sciences 55B: 69-84.

Ackerman, P. L., 2003: Cognitive ability and non-ability trait determinants of expertise. Educational Researcher 32 (8): 15-20.

Ackerman, P. L., 2008: Knowledge and cognitive aging. S. 445-489 in: F. I. M. Craik und T. A. Salthouse (Hg.): The handbook of aging and cognition (3rd ed.). New York, NY, USA: Psychology Press.

Ackerman, P. L. und M. E. Beier, 2004: Knowledge and Intelligence. S. 125-139 in: O. Wilhelm und R. Engle (Hg.): Handbook of understanding and measuring intelligence. Thousand Oaks, CA: Sage.

Ackerman, P. L., K. R. Bowen, M. B. Beier und R. Kanfer, 2001: Determinants of individual differences and gender differences in knowledge. Journal of Educational Psychology 93: 797-825.

Ackerman, P. L. und E. D. Heggestad, 1997: Intelligence, personality, and interests: Evidence for overlapping traits. Psychological Bulletin 121: 219-245.

Ackerman, P. L. und E. L. Rolfhus, 1999: The locus of adult intelligence: Knowledge, abilities, and non-ability traits. Psychology and Aging 14: 314–330.
Baltes, P. B., 1987: Theoretical propositions of life-span developmental psychology: on the dynamics between growth and decline. Developmental Psychology 23: 611-626.
Baltes, P. B., U. M. Staudinger und U. Lindenberger, 1999: Lifespan Psychology: Theory and Application to Intellectual Functioning. Annual Review of Psychology 50: 471-507.
Barrick, M. R., M. K. Mount und T. A. Judge, 2001: Personality and performance at the beginning of the new millennium: What do we know and where do we go next? International Journal of Selection and Assessment 9: 9-20.
Binet, A. und T. Simon, 1905: Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. L'année psychologique 11: 191-244.
Brennan, R. L., 2001: Generalizability Theory. New York: Springer.
Carroll, J. B., 1993: Human cognitive abilities: A survey of factor-analytic studies. New York: Cambridge University Press.
Carroll, J. B., 2003: The higher stratum structure of cognitive abilities: Current evidence supports g and about ten broad factors. S. 153-193 in: H. Nyborg (Hg.): The scientific study of general intelligence: Tribute to Arthur R. Jensen. New York: Pergamon.
Cattell, R. B., 1943: The measurement of adult intelligence. Psychological Bulletin 40: 153-193.
Cattell, R.B., 1963: Theory of fluid and crystallized intelligence: A critical experiment. Jounal of Educational Psychology 54: 1-22.
Cattell, R. B., 1971: Abilities: Their Structure, Growth, and Action. Boston, MA: Houghton Mifflin.
Ceci, S. J., 1991: How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. Developmental Psychology 27: 703–722.
Cliffordson, C. und J.-E. Gustafsson, 2008: Effects of age and schooling on intellectual performance: Estimates obtained from analysis of continuous variation in age and length of schooling. Intelligence 36: 143-152.
Cohen, J., 1988: Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale: Lawrence Erlbaum Associates.
Collins, L. M., J. L. Schafer und C.-M. Kam, 2001: A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods 6: 330-351.
Cronbach, L. J., 1949: Essentials of psychological testing. New York: Harper.
Danthiir, V., R. D. Roberts, R. Schulze und O. Wilhelm, 2004: Mental Speed: On Frameworks, Paradigms, and a Platform for the Future. S. 27-46 in: O. Wilhelm und R. W. Engle (Hg.): Handbook of Understanding and Measuring Intelligence. London: Sage.
Ebbinghaus, H., 1897: Über eine neue Methode zur Prüfung geistiger Fähigkeiten und ihre Anwendung bei Schulkindern. Zeitschrift für Psychologie und Physiologie der Sinnesorgane 13: 401-459.
Ehmke, T. und T. Siegle, 2005: ISEI, ISCED, HOMEPOS, ESCS. Indikatoren der sozialen Herkunft bei der Quantifizierung von sozialen Disparitäten. Zeitschrift für Erziehungswissenschaft 8: 521-539.
Furnham, A. und G. Dissou, 2007: The relationship between self-estimated and test-derived scores of personality and intelligence. Journal of Individual Differences 28: 37-44.
Ganzeboom, H. B. G., P. M. De Graaf und D. J. Treiman, 1992: A Standard International Socio-Economic Index of Occupational Status. Social Science Research 21: 1-56.
Grabner, R. H. und E. Stern, 2010: Measuring Cognitive Ability. S. 753-768 in: Rat für Sozial- und Wirtschaftsdaten (Hg.): Building on progress: Expanding the research infrastructure for the social, economic, and behavioral sciences. Opladen: Budrich UniPress.

Graham, J. W., A. E. Olchowski und T. D. Gilreath, 2007: How many imputations are really needed? Some Practical Clarifications of Multiple Imputation Theory. Prevention Science 8: 206-213.

Hebb, D. O., 1942: The effect of early and late brain injury upon test scores and the nature of normal adult intelligence. Proceedings of the American Philosophical Society 85: 275-292.

Horn, J. L., 1965: Fluid and crystallized intelligence: A factor analytic study of the structure among primary mental abilities. Dissertation, University of Illinois.

Horn, J. L., 1988: Thinking about human abilities. S. 645-685 in: J. R. Nesselroade (Hg.): Handbook of multivariate psychology. New York: Academic Press.

Horn, J. L., und R. B. Cattell, 1966: Refinement and test of the theory of fluid and crystallized general intelligences. Journal of Educational Psychology 57: 253-270.

Horn, J. L. und J. Noll, 1997: Human cognitive capabilities: Gf–Gc theory. S. 53-91 in: D. P. Flanagan, J. L. Genshaft und P. L. Harrison (Hg.): Contemporary intellectual assessment: Theories, tests and issues. New York: Guilford.

Hülür, G., O. Wilhelm und S. Schipolowski, 2011: Prediction of self-reported knowledge with over-claiming, fluid and crystallized intelligence and typical intellectual engagement. Learning and Individual Differences 21: 742-746.

ILO, 2007: International Standard Classification of Occupations ISCO-08. http://www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm (26.11.2012).

Lang, F. R., D. Weiss, A. Stocker und B. von Rosenbladt, 2007: Assessing Cognitive Capacities in Computer-Assisted Survey Research: Two Ultra-Short Tests of Intellectual Ability in the German Socio-Economic Panel (SOEP). Schmollers Jahrbuch 127: 183-192.

Li, S.-C, 2003: Biocultural orchestration of developmental plasticity across levels: The interplay of biology and culture in shaping the mind and behavior across the life span. Psychological Bulletin 129: 171-194.

Lüdtke, O., A. Robitzsch, U. Trautwein und O. Köller, 2007: Umgang mit fehlenden Werten in der psychologischen Forschung: Probleme und Lösungen. Psychologische Rundschau 58: 103-117.

Lynn, R. und P. Irwing, 2002: Sex differences in general knowledge, semantic knowledge and reasoning ability. British Journal of Psychology 93: 545-556.

Lynn, R., P. Irwing und T. Cammock, 2001: Sex differences in general knowledge. Intelligence 30: 27-40.

Lynn, R., S. Wilberg und J. Margraf-Stiksrud, 2004: Sex differences in general knowledge in German high school students. Personality and Individual Differences 37: 1643-1650.

Mead, A. D. und F. Drasgow, 1993: Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. Psychological Bulletin 114: 449–458.

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

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: 1-10.

Muthén, L. K. und B. O. Muthén, 1998-2010: Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Neumann, M., G. Nagy, U. Trautwein und O. Lüdtke, 2009: Vergleichbarkeit von Abiturleistungen: Leistungs- und Bewertungsunterschiede zwischen Hamburger und Baden-Württemberger Abiturienten und die Rolle zentraler Abiturprüfungen. Zeitschrift für Erziehungswissenschaft 12: 691-714.

Ostendorf, F. und A. Angleitner, 2004: NEO-Persönlichkeitsinventar nach Costa und McCrae, Revidierte Fassung (NEO-PI-R). Göttingen: Hogrefe.

Paulhus, D. L. und P. D. Harms, 2004: Measuring cognitive ability with the overclaiming technique. Intelligence 32: 297-314.

Rammstedt, B., C. J. Kemper, M. C. Klein, C. Beierlein und A. Kovaleva, 2012: Eine kurze Skala zur Messung der fünf Dimensionen der Persönlichkeit: Big-Five-Inventory-10 (BFI-10). GESIS-Working Papers 2012|23. Köln: GESIS.

Rammstedt, B. und O. P. John, 2007: Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of Research in Personality 41: 203-212.

Rat für Sozial- und Wirtschaftsdaten, 2010: Building on progress: Expanding the research infrastructure for the social, economic, and behavioral sciences. Opladen: Budrich UniPress.

Raykov, T., D. M. Dimitrov und T. Asparouhov, 2010: Evaluation of scale reliability with binary measures using latent variable modeling. Structural Equation Modeling: A Multidisciplinary Journal 17: 265–279.

Reise, S. P., T. M. Moore und M. G. Haviland, 2010: Bifactor Models and Rotations: Exploring the Extent to Which Multidimensional Data Yield Univocal Scale Scores. Journal of Personality Assessment 92: 544–559.

Rhemtulla, M. und T. D. Little, 2012: Planned Missing Data Designs for Research in Cognitive Development. Journal of Cognition and Development 13: 425-438.

Rolfhus, E. L. und P. L. Ackerman, 1999: Assessing individual differences in knowledge: Knowledge structures and traits. Journal of Educational Psychology 91: 511-526.

Rowe, D. C., K. C. Jacobson und E. J. van den Oord, 1999: Genetic and environmental influences on vocabulary IQ: parental education level as moderator. Child Development 70: 1151–1162.

Rubin, D. B., 1976: Inference and missing data. Biometrika 63: 581-592.

Schipolowski, S., U. Schroeders und O. Wilhelm, 2008: BEFKI - Berlin Test of Fluid and Crystallized Intelligence. Präsentation auf dem XXIX International Congress of Psychology, Berlin.

Schroeders, U. und O. Wilhelm, 2010: Testing reasoning ability with handheld computers, notebooks, and paper and pencil. European Journal of Psychological Assessment 26: 284–292.

Schroeders, U., S. Schipolowski, C. Nelles und O. Wilhelm, 2011: Interest, domain specific knowledge, and fluid intelligence: Profile covariances and prediction of vocational training success. Präsentation auf dem 15th Biennial Meeting of the International Society for the Study of Individual Differences, London, UK.

Schroeders, U., S. Schipolowski und O. Wilhelm, 2010: Berliner Test zur Erfassung fluider und kristalliner Intelligenz. Präsentation auf dem 47. Kongress der Deutschen Gesellschaft für Psychologie, Bremen.

Schulze, R., 2004: Modeling Structures of Intelligence. S. 241-263 in O. Wilhelm und R. Engle (Hg.): Handbook of understanding and measuring intelligence. Thousand Oaks, CA: Sage.

Sinharay, S., 2010: How Often Do Subscores Have Added Value? Results from Operational and Simulated Data. Journal of Educational Measurement 47: 150-174.

Sinharay, S., S. Haberman und G. Puhan, 2007: Subscores Based on Classical Test Theory: To Report or Not to Report. Educational Measurement: Issues and Practice 26 (4): 21-28.

Spearman, C. E., 1938: Measurement of intelligence. Scientia 64: 75-82.

UNESCO, 1997: International Standard Classification of Education ISCED 1997. http://www.unesco.org/education/information/nfsunesco/doc/isced_1997.htm (20.11.2012).

Widaman, K. F., T. D. Little, K. J. Preacher und G. M. Sawalani, 2011: On creating and using short forms of scales in secondary research. S. 39-61 in: K. H. Trzesniewski, M. B. Donnellan und R. E. Lucas (Hg.): Secondary Data Analysis: An Introduction for Psychologists. Washington, DC: American Psychological Association.

Wilhelm, O. und S. Schipolowski, 2010: Intelligenzdiagnostik in der Pädagogischen Psychologie. In G. L. Huber (Hg.): Enzyklopädie Erziehungswissenschaft Online. Fachgebiet Pädagogische Psychologie. Weinheim/München: Juventa.

Wilhelm, O., U. Schroeders und S. Schipolowski, in Vorbereitung: BEFKI – Berliner Test zur Erfassung fluider und kristalliner Intelligenz. Mittelstufenform. Göttingen: Hogrefe.

Wolf, C., 1995: Sozio-ökonomischer Status und berufliches Prestige. ZUMA-Nachrichten 37: 102-136.

Von der Heyde, C., 2009: Das ADM-Stichprobensystem für persönlich-mündliche Befragungen. http://www.adm-ev.de/fileadmin/user_upload/PDFS/Beschreibung-ADM-Stichproben-f2f_DE.pdf (30 KB) (20.11.2012)

Von Stumm, S. und P. L. Ackerman, 2012: Investment and Intellect: A Review and Meta-Analysis. Psychological Bulletin: Advance online publication, doi: 10.1037/a0030746.

Yu, C. Y., 2002: Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. Doctoral dissertation, University of California, Los Angeles.

Ziegler, M., E. Danay, M. Heene, J. Asendorpf und M. Bühner, 2012: Openness, fluid intelligence, and crystallized intelligence: Toward an integrative model. Journal of Research in Personality 46: 173-183.

Ziegler, M., C. J. Kemper und B. Rammstedt, 2013: The Vocabulary and Overclaiming Test (VOC-T). Journal of Individual Differences 34: 32-40.

Ziegler, M., C. MacCann und R. D. Roberts, 2011: New Perspectives on Faking in Personality Assessment. New York: Oxford University Press.

Zumbo, B. D., A. M. Gadermann und C. Zeisser, 2007: Ordinal versions of coefficients alpha and theta for Likert rating scales. Journal of Modern Applied Statistical Methods 6: 21-29.

Anschrift des Autors

Stefan Schipolowski
Humboldt-Universität zu Berlin
Institut zur Qualitätsentwicklung im Bildungswesen
Unter den Linden 6
10099 Berlin
stefan.schipolowski@iqb.hu-berlin.de

Ko-Autor/-innen

Oliver Wilhelm
Institut für Psychologie und Pädagogik
Universität Ulm

Ulrich Schroeders
Institut zur Qualitätsentwicklung im Bildungswesen
Humboldt-Universität zu Berlin

Anastassiya Kovaleva
Fakultät für Biologie, Universität Bielefeld

Christoph J. Kemper
Institut für medizinische und pharmazeutische Prüfungsfragen, Mainz

Beatrice Rammstedt
GESIS – Leibniz-Institut für Sozialwissenschaften, Mannheim

methoden, daten, analysen · 2013, Jg. 7(2), S. 183-208 DOI: 10.12758/mda.2013.011

# Die Kurzform des Hagener Matrizen-Tests (HMT-S)

## *Ein 6-Item Intelligenztest zum schlussfolgernden Denken*

# The Short Version of the Hagen Matrices Test (HMT-S)

## *A 6-Item Induction Intelligence Test*

*Timo Heydasch, Julia Haubrich, Karl-Heinz Renner*

*Zusammenfassung*
Die Kurzform des Hagener Matrizen-Tests (HMT-S) ist ein kostenfreier Online-Intelligenztest, der bezogen auf das CHC- Modell der Intelligenz (Schneider/McGrew 2012) Induktion bzw. Reasoning, die Fähigkeit zum schlussfolgernden Denken misst. Der 6-Item HMT-S basiert auf dem 20 Items umfassenden Hagener Matrizen-Test (HMT; Heydasch/Renner/Haubrich/Hilbig/Zettler 2013). Die interne Konsistenz des HMT-S betrug in den durchgeführten Studien .62. Die Korrelationen des HMT-S zur Langversion betrug in einer ersten Studie $r$ = .79 und $r$ = .78 in einer zweiten. Zudem konnte die konvergente Validität durch Zusammenhänge u.a. zum Intelligenz-Struktur-Test 2000 R (Liepmann/Beauducel/Brocke/Amthauer 2007) belegt werden. Die Kriteriumsvalidität wurde durch Assoziationen zu akademischen Erfolgen nachgewiesen. Der HMT-S ist somit ein reliabler, valider und ökonomischer Intelligenztest, dessen Einsatz über folgende URL angefordert werden kann: http://HMT.de.lv

*Abstract*
The short version of the Hagen Matrices Test (HMT-S) is a free of charge online intelligence test measuring induction in reference to the CHC model of intelligence (Schneider/McGrew 2012). The 6-item HMT-S is based on the 20-item Hagen Matrices Test (HMT; Heydasch/Renner/Haubrich/Hilbig/Zettler 2013). The internal consistency of the HMT-S in our studies was .62. The correlations with the original scale were $r$ = .79 in a first study and $r$ = .78 in a second one. In addition, convergent validity was shown by correlations with the Intelligence Structure Test 2000 R (Liepmann/Beauducel/Brocke/Amthauer 2007). Associations with academic performance indicated criterion related validity. To sum it up, the HMT-S is a reliable, valid, economic, and efficient intelligence test. Free applications can be requested via http://HMT.de.lv

# 1 Einleitung

Bedeutende Bereiche des Lebens stehen mit Intelligenz in Zusammenhang. Sowohl für den Erfolg in der Schule oder im Studium (Rindermann/Neubauer 2000; Kuncel/Hezlett/Ones 2004; Kunina/Wilhelm/Formazin/Jonkmann/Schroeders 2007; Poropat 2009) als auch im beruflichen Kontext (Harrell/Harrell 1945; Hunter/Hunter 1984; Schmidt/Hunter 1998; Judge/Higgins/Thoresen/Barrick 1999; Salgado/Anderson/Moscoso/Bertua/de Fruyt/Rolland 2003; Kuncel/Hezlett/Ones 2004; Ng/Eby/Sorensen/Feldman 2005; Hülsheger/Maier/Stumpp 2007; Ziegler/Dietl/Danay/Vogel/Bühner 2011; vgl. auch Strenze 2007) spielt Intelligenz eine zentrale Rolle. Aber auch andere Bereiche wie psychische Gesundheit (z.B. Khandaker/Barnett/White/Jones 2011), Ehescheidungen und die Lebensdauer (Roberts/Kuncel/Shiner/Caspi/Goldberg 2007) sind mit Intelligenz assoziiert.

Ist bei einer Studie die Berücksichtigung der Intelligenz angezeigt – sei es als Prädiktor, Moderator, Kontrollvariable (vgl. Blickle/Kramer/Mierke 2010), Mediator oder Kriterium – so stellt sich anscheinend zunächst die Qual der Wahl, aus dem vielfältigen Angebot von Intelligenztestverfahren (vgl. Leibniz-Zentrum für Psychologische Information und Dokumentation 2012) ein passendes Instrument auszuwählen. Neben Überlegungen zur Intelligenzdefinition, zum zugrundeliegenden Intelligenzmodell und der Festlegung des spezifischen Intelligenzbereiches spielen ökonomische Erwägungen eine Rolle. So können u.a. zwei Probleme resultieren, die dem Einsatz eines spezifischen Verfahrens entgegenstehen: 1. Die Instrumente sind zu lang und erfordern ein nicht zu realisierendes Maß an Testadministratoren- oder Teilnehmerzeit und/oder 2. handelt es sich um kommerzielle Verfahren, die teilweise zugangsbeschränkt über Verlage erworben werden müssen und für deren Einsatz ggf. erhebliche Kosten entstehen.

Kostenfreie Kurztests zur Messung der allgemeinen Intelligenz oder verschiedener Intelligenzbereiche sind Mangelware. Der 20 Items umfassende Hagener Matrizen-Test (HMT; Heydasch/Renner/Haubrich/Hilbig/Zettler 2013) stellt als nicht-kommerzielles Online-Verfahren eine mögliche Option zur Intelligenzmessung dar. Sowohl Reliabilität (interne Konsistenz und Retestreliabilität) als auch konvergente und divergente Validität wurden belegt. Zudem sprechen Zusammenhänge zur Schulausbildung und zu universitären Leistungen für dessen Kriteriumsvalidität. Kritisch anzumerken ist jedoch, dass der HMT mit 20 Aufgaben womöglich für einige Einsatzbereiche zu zeitintensiv ist. Dies gilt insbesondere für Bereiche, in denen Kosteneinsparungen durch kurze und somit schnelle Verfahren zwingend sind bzw. die Akzeptanz von Teilnehmern und die Bereitschaft zur Teilnahme möglichst hoch sein sollen. Ein anderer Kritikpunkt bezieht sich auf die

Schwierigkeit. Da der HMT primär zur Studienerfolgsprädiktion entwickelt wurde, sind eher schwierige Matrizen enthalten, damit eine angemessene Differenzierung bei Studierenden und Studiumsinteressierten erzielt werden kann. Aus diesen Gründen war die Entwicklung eines kürzeren und weniger schwierigen Intelligenztests erforderlich. Umgesetzt wurde dies, indem wir auf Basis des HMT die Kurzversion des Hagener Matrizen-Tests, den HMT-S, entwickelten und somit nun ein weiterer kostenfreier Intelligenztest vorliegt.

Konsens über das, was Intelligenztests messen bzw. über eine Definition des Intelligenzbegriffes existieren aber nicht (Willis/Dumont/Kaufman 2011; Wasserman 2012). Im Bestreben, ein gewisses Maß an Übereinkunft bezüglich einer Definition herzustellen, unterzeichneten 1994 52 Experten (u.a. Carroll, Cattell, Eysenck, Horn, Jensen, Thorndike und Vernon) eine Definition, die sie als Mainstream in der Intelligenzforschung akzeptierten (Gottfredson 1997: 13):

> Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly, and learn from experience. It is not merely book-learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings- "catching on", "making sense" of things, or "figuring out" what to do.

Unter anderem auf dieser Ausführung basierend definiert Sternberg (2005: 751) Intelligenz als „*the capacity to learn from experience, using metacognitive processes to enhance learning, and the ability to adapt to the surrounding environment, which may require different adaptations within different social and cultural contexts*". Intelligente Personen sind also lernfähig und können sich an den jeweiligen Kontext anpassen und sind daher schließlich erfolgreich.

Bezüglich einer Theorie der Intelligenz oder dem hierarchischen Aufbau der kognitiven Fähigkeitskomponenten besteht ebenfalls kein Einvernehmen (Gottfredson/Saklofske 2009); vielmehr liegt eine Vielzahl an Theorien und Modellen vor (vgl. Davidson/Kemp 2011). Das Cattell-Horn-Carroll(CHC)-Modell der Intelligenz (Schneider/McGrew 2012; vgl. auch McGrew 1997; McGrew 2005) ist aber eine Taxonomie, die die theoretische Einordnung und Integration von Intelligenztheorien und Intelligenztests wie dem HMT-S ermöglicht. Das CHC-Modell basiert auf der Horn-Cattell Gf-Gc Theorie (Horn/Noll 1997; Horn/Blankson 2012) und der Three-Stratum-Theory (Carroll 1993; 2005) und ordnet die Intelligenzbereiche auf drei hierarchischen Ebenen an. An der Spitze steht *g*, die allgemeine Intelligenz (vgl. Spearman 1904; Jensen 1998). Die Ebene unterhalb von *g* enthält 16 breite Fähigkeitsbereiche, darunter *fluid reasoning* (Gf), *short-term memory* (Gsm), *processing speed* (Gs), *psychomotor speed* (Gps), *domain-specific knowledge* (Gkn) und *visual processing* (Gv). Unterhalb dieser Ebene werden verschiedene spezifi-

sche Fähigkeitsbereiche benannt, die als Facetten des jeweiligen breiten Fähigkeitsbereichs zu verstehen sind. Innerhalb von Gf werden beispielsweise die Bereiche *induction* (I), *general sequential reasoning* (RG, auch Deduktion genannt; vgl. Evans 2005) und *quantitative reasoning* (RQ) als gut gesicherte Facetten aufgeführt.

Ausgehend vom Aufgabenmaterial – den Matrizen – und der theoretischen Einordnung des HMT (Heydasch et al. 2013), erwarten wir, dass der HMT-S Induktion, *„the ability to observe a phenomenon and discover the underlying principles or rules…"* (Schneider/McGrew 2012, S. 112; vgl. auch Sloman/Lagnado 2005), misst. Dabei operationalisiert der HMT-S Gf, denn: *„Induction is probably the core aspect of Gf"* (Schneider/McGrew 2012, S. 112; siehe auch Carroll 1993). Zwischen Gf und *g* existiert wiederum ein Zusammenhang: Schneider und McGrew führen an, dass Cattell *g* als lebenslang kumuliertes Gf auslegte und dass andere Autoren Gf und *g* sogar als identisch betrachten.

In der Intelligenzforschung wurden von einigen Autoren alternative Modelle vorgestellt, die kognitive Fähigkeiten bzw. Aufgabenmaterial (neben anderen Merkmalen) nach Inhaltsbereichen differenzieren (vgl. Süß/Beauducel 2004). Dazu gehören das Radex-Modell (Guttman 1965; Guttman/Levy 1991), das Intelligenzstrukturmodell (Guilford 1967), das Berliner Intelligenzstrukturmodell (BIS; Jäger 1982) und das Hierarchische Rahmen- bzw. Protomodell der Intelligenzstrukturforschung (HPI; Amthauer/Brocke/Liepmann/Beauducel 2001; Liepmann/Beauducel/Brocke/Amthauer 2007). Diesen Modellen ist gemeinsam, dass verbale, numerische und figural-(bildhafte) Fähigkeiten bzw. Qualitäten des Aufgabenmaterials unterschieden werden. So werden figurale Matrizenaufgaben in der Regel dem figuralen Intelligenzbereich zugeordnet und demnach wäre der HMT-S als ein figuraler Induktionstest zu klassifizieren.

Die Validität des HMT-S sollte trotz der Testkürzung gegeben und zum HMT vergleichbar sein. Die Validität eines Kurztests ist zunächst als Übereinstimmung mit der Langversion definiert (Silverstein 1990). Da individualdiagnostische Erwägungen beim HMT-S als Forschungsinstrument nicht ins Gewicht fallen (z.B. individuelle Schätzung des Intelligenzquotienten oder Hochbegabungsdiagnostik), ist die Validität als Korrelation zwischen den Testversionen zu verstehen.

Aus der Korrelation bzw. der gemeinsamen Varianz von HMT-S und HMT kann nicht zwangsläufig auf die Konstrukt- oder Kriteriumsvalidität der Kurzversion geschlossen werden, sodass weitere Kennwerte in diesen Bereichen erforderlich sind: Es sollten sich hohe Validitätskoeffizienten zum einen zu Erfolgsindikatoren in Schule, Studium und Beruf ergeben und zum anderen zu Messwerten, die als Indikatoren schlussfolgernden Denkens allgemein akzeptiert sind. Validitätskoeffizienten sollten geringer ausfallen bei nur teilweise überlappenden Konstrukten.

Die Ähnlichkeit teilweise überlappender Konstrukte nimmt ab, wenn andere Intelligenzbereiche (z.B. Gkn oder Gsm versus Gf), unterschiedlich abstrakte Intelligenzbereiche (z.B. g versus I) oder andere Inhaltsbereiche (verbal oder numerisch versus figural) zum Vergleich herangezogen werden. Ebenfalls ist anzunehmen, dass durch unterschiedliche Datenquellen (z.B. Selbsteinschätzung versus objektiver Leistungstest, vgl. Cattell 1957; Pawlik 2006) oder durch verschiedene Erhebungsmethoden (z.B. Online-Test versus Paper-Pencil-Test; vgl. Guttman 1965; Guttman/Levy 1991) methodenbedingte Varianz die Korrelationen theoretisch nah stehender Konstrukte mindert. Unabhängigkeit sollte sich erweisen zu theoretisch nicht assoziierten Bereichen.

## 2 Methode

Zur Entwicklung und Validierung des HMT-S wurde die Stichprobe genutzt, die der Validierung des HMT diente (Heydasch et al. 2013). Diese Stichprobe wurde unter Verwendung der SPSS Statistiksoftware (Version 21) per Zufall in zwei annähernd gleich große Gruppen geteilt. Auf Basis der Datenanalyse der ersten Gruppe (Studie 1) wurden sechs Items aufgrund ihrer Schwierigkeiten ($p > .20$), Trennschärfen ($r_{it} > .30$) und Korrelationen mit dem Intelligenz-Struktur-Test 2000 R (Liepmann et al. 2007; $r > .30$) selektiert. Die resultierenden Kennwerte der nachträglich berechneten Kurzversion in Studie 1 wurden mit der zweiten Gruppe kreuzvalidiert (Studie 2). Zur weiteren Bestätigung der Kennwerte folgte eine dritte Studie, in der der HMT-S schließlich in der finalen Fassung vorgelegt wurde.

### 2.1 Stichprobe

In allen drei Studien bestanden die Stichproben aus Studierenden im Studiengang Psychologie (Bachelor of Science) der FernUniversität in Hagen, die für die Teilnahme Versuchspersonenstunden erhielten. Über alle drei Studien hinweg haben $N = 1.572$ Probanden an den Testungen und Befragungen teilgenommen. Der Anteil der Frauen betrug 75% und das Alter lag bei $M = 31.6$ Jahren ($SD = 8.97$). In Tabelle 1 werden die Stichprobengrößen, die Kennwerte für das Geschlechterverhältnis und die Altersverteilungen für die einzelnen Studien aufgeführt.

Tabelle 1 Stichprobengrößen, Anteile weiblicher Probandinnen und Kennwerte der Altersverteilung

| | | | Alter | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Percentile | | | | | | |
| Studie | *N* | Frauen | 10% | 25% | 50% | 75% | 90% | *M* | *SD* |
| 1 | 681 | 74% | 21 | 24 | 30 | 37 | 45 | 31.5 | 8.89 |
| 2 | 658 | 74% | 22 | 25 | 30 | 38 | 45 | 31.8 | 8.95 |
| 3 | 233 | 80% | 21 | 24 | 30 | 39 | 46 | 31.6 | 9.28 |

## 2.2 Instrumente

Der HMT und der daraus entwickelte HMT-S gliedern sich in den Instruktionsteil, den Aufgabenteil und den Schlussteil. In der Instruktion wird das Aufgabenprinzip erläutert: Die Probanden sollen den regelgeleiteten Aufbau der unvollständigen figuralen 3x3 Matrizen[1] erkennen und aus einer Auswahl von acht Lösungsalternativen die korrekte identifizieren und markieren. Dazu stehen zwei Minuten pro Aufgabe zur Verfügung, wobei die Lösung auch schon vor Ablauf der Zeit abgesendet werden kann. Die Regeln, nach denen die Matrizen aufgebaut sein können, werden vorab benannt: Addition, Subtraktion sowie räumliche Verschiebungen (Elemente „bewegen" sich von einer Seite zur anderen, rotieren etc.). Das Aufgabenprinzip wird dabei durch zwei Beispielaufgaben verdeutlicht[2]. Der Aufgabenteil besteht beim HMT aus 20 bzw. beim HMT-S aus sechs Matrizen. Falls ein Testteilnehmer nicht über den „Weiter"-Button vor Ablauf der zweiminütigen Frist eine Lösung absendet, wird automatisch die ggf. markierte Lösung registriert und die folgende Aufgabe präsentiert. Zur zeitlichen Orientierung wird eine Uhr eingeblendet. Die Auswertung erfolgt online während der Testbearbeitung. Korrekte Lösungen werden mit 1, inkorrekte oder nicht bearbeitete Matrizen (auch Aufgaben, die nach einem Abbruch nicht bearbeitet wurden) werden mit 0 codiert und anschließend summiert. Im Schlussteil werden den Teilnehmern das individuelle Ergebnis in Form von Anzahl und Prozentsatz korrekt gelöster Aufgaben mitgeteilt. Zusätzlich wird der HMT bzw. der HMT-S für Laien verständlich inhaltlich näher skizziert und beispielsweise darauf hingewiesen, dass der Test nicht perfekt reliabel ist oder Zusammenhänge zu akademischen Erfolgen nicht deterministisch zu verstehen sind. Auf

1 Matrizen freundlich überlassen von Lutz Hornke.

2 Die Instruktion mit den Beispielaufgaben ist online einzusehen unter der URL https://ww3.unipark.de/uc/HMT_S_Vorschau/

die Berechnung und Rückmeldung eines IQs wird beim HMT-S im Gegensatz zum HMT verzichtet.

Zur weiteren Validierung wurden folgende Instrumente herangezogen:

- *Intelligenz-Struktur-Test 2000 R* (I-S-T 2000 R; Liepmann et al. 2007; vgl. Schmidt-Atzert 2002; Schmidt-Atzert/Rauch 2008): Der I-S-T 2000 R erfasst als objektiver Leistungstest folgende Intelligenzbereiche: Schlussfolgendes Denken (Reasoning, R), Wissen (W) und Merkfähigkeit (ME). R und W werden jeweils gebildet aus Skalen mit verbalen, numerischen und figuralen Inhalten, die wiederum jeweils drei Aufgabengruppen enthalten. ME basiert auf einer verbalen und einer figuralen Aufgabengruppe. Weiterhin lassen sich die als unabhängig konzipierten Faktoren für fluide ($g_f$) und kristalline ($g_c$) Intelligenz bestimmen. Der I-S-T 2000 R bezieht sich theoretisch auf das Radex-Modell (Guttman 1965; Guttman/Levy 1991) bzw. das BIS (Jäger, 1982), indem Aufgabeninhalte zu unterscheiden sind nach verbaler, numerischer und figuraler Qualität. Gleichzeitig beziehen sich die Autoren mit der Konzeption der Dimensionen R und W auf Cattell (1987). Die beiden Ansätze wurden von Amthauer und Kollegen (2001) zum HPI synthetisiert, das schließlich die theoretische Grundlage des I-S-T 2000 R darstellt.
- *Inventar zur selbsteingeschätzten Intelligenz* (ISI; Rammstedt/Rammsayer 2002): Das ISI misst in einer Online-Adaption basierend auf dem Konzept multipler Intelligenzen (Gardner 1983) die selbsteingeschätzten Intelligenzen in den Bereichen verbales Verständnis, Wortflüssigkeit, mathematische Intelligenz, räumliche Intelligenz, Gedächtnisfähigkeit, Wahrnehmungsgeschwindigkeit, logisches Denken, musikalische Intelligenz, körperlich-kinästhetische Intelligenz, interpersonale Intelligenz sowie intrapersonale Intelligenz.
- *Allgemeine Selbstwirksamkeitserwartung* (SWE; Schwarzer/Jerusalem 1995): Mit der Skala SWE wird die situationsübergreifende Überzeugung erfasst, mit Problemen und Herausforderungen umgehen und diese erfolgreich bewältigen zu können.
- *Studiumspezifische Selbstwirksamkeit* (Schiefele/Moschner/Husstegge 2002): In Anlehnung an die Skala zur Messung der Allgemeinen Selbstwirksamkeitserwartung misst die für diese Untersuchung an das Fernstudium adaptierte Skala die Überzeugung, erfolgreich mit Herausforderungen im Studienkontext umgehen zu können.
- *Allgemeine Hilflosigkeit* (Jerusalem/Schwarzer 1986, 2012): Auf Basis der Theorie der gelernten Hilflosigkeit von Seligman (1975) erfasst diese Skala bereichsübergreifend die selbsteingeschätzte Hilflosigkeit bei der Bewältigung von Problemen und Herausforderungen.

- *Studiumspezifische Hilflosigkeit* (Jerusalem/Schwarzer 1986, 2012): Mit diesem Fragebogen wird die Erwartung von Studierenden gemessen, Probleme im Kontext eines Studiums nicht meistern zu können.
- *Instrumente zum Selbstkonzept*: Die drei Skalen *generelles akademisches Selbstkonzept, mathematisches Selbstkonzept* und *sprachliches Selbstkonzept* (Schiefele et al. 2002) erfassen für die aufgeführten Bereiche das Selbstverständnis einer Person, das sich auf vorhandene Fähigkeiten und Problemlösekompetenzen bezieht.
- *Revised Achievement Motivation Scale* (AMS-R; Lang/Fries 2006): Die AMS-R misst das explizite Leistungsmotiv mittels der Dimensionen *hope of success* und *fear of failure,* die die Annäherungs- bzw. die Vermeidungstendenz des Leistungsmotivs abbilden.

Zusätzlich wurden demographische Variablen (Geschlecht und Alter) erhoben, das Land des Schulabschlusses und der erreichte Schulabschluss (codiert mit *Allgemeine Hochschulreife* = 3, *Fachhochschulreife oder fachgebundene Hochschulreife* = 2, *Mittlerer Schulabschluss* = 1). Zudem berichteten die Teilnehmer ihre Schulabschlussnote und jeweils die letzte erreichte Note in den Fächern Mathematik, Englisch, Deutsch, Biologie, Kunst und Sport. Ergänzend gaben die Teilnehmer ihre Modulprüfungsnoten im Studiengang B.Sc. Psychologie an. Zur Berechnung eines interindividuell vergleichbaren Studienerfolgsindikators wurden die Modulabschlussnoten zunächst standardisiert und anschließend gemittelt, um für Niveauunterschiede zwischen den Modulprüfungen zu kontrollieren. Für die Berechnung des Indikators des Studienerfolgs musste mindestens eine Note angegeben worden sein.

## 2.3 Durchführung

Die Datenerhebung erfolgte weitestgehend online und unbeaufsichtigt mit EFS-Survey der QuestBack GmbH (vgl. Buchwald/Spoden/Fleischer/Leutner 2013). Die Probanden haben so sowohl den Ort (z.B. heimischer PC oder mobiler Laptop) als auch den Zeitpunkt der Teilnahme selbst bestimmt. Auf einer Web-Seite (http://www.fernuni-hagen.de/psychologie/forschung/vlabor.shtml) konnten sich die potentiellen Teilnehmer über die Teilnahmebedingungen der universitären Online-Studien informieren und aus verschiedenen zeitgleich angebotenen Projekten wählen. Nach Zugang zu einem Projekt wurden die Inhalte und Teilnahmebedingungen der jeweiligen Studie spezifiziert und die Teilnehmer erklärten anschließend ihr Einverständnis zur Teilnahme.

Bei den Hagener Matrizen-Tests wurden zu Beginn die grundlegenden demographischen Daten (z.B. Alter und Geschlecht) und darauf folgend die selbsteingeschätzten Intelligenzen mit dem ISI erhoben. Im Anschluss wurde die jeweilige Version des Hagener Matrizen-Tests vorgelegt. Zuletzt wurden den Teilnehmern die individuellen Ergebnisse des HMT bzw. des HMT-S zurückgemeldet. Aufgrund der großen Anzahl und teilweise enormen Länge der vorgelegten Instrumente[3] wurden bei Studie 1 und 2 die Verfahren zur Messung der Selbstwirksamkeit, zur Hilflosigkeit, zum Selbstkonzept und zum Leistungsmotiv nicht im gleichen, sondern in unabhängigen Online-Projekten integriert. In Studie 3 wurden diese Instrumente trotz der resultierenden Länge in das Projekt des HMT-S integriert und zwischen den demographischen Variablen und dem ISI positioniert.

Der I-S-T 2000 R wurde abweichend zu den Online-Erhebungen als Paper-Pencil-Verfahren durchgeführt[4]. Studierende haben sich dafür für vorgegebene Termine und Orte registriert. Die Testungen wurden anschließend entsprechend den Anweisungen im Manual umgesetzt.

Bei jeder Sitzung, in der Fragebögen oder Tests bearbeitet wurden, erstellten die Teilnehmer regelgeleitet einen sechsstelligen individuellen Pseudonymisierungscode, der immer nach denselben Vorgaben gebildet wurde. So wurde jeder Person ein bestimmter Code zugeordnet. Anhand der Psyeudonymisierungscodes konnten die Daten verschiedener Sitzungen einer Person zusammengeführt werden, ohne die Identität der Teilnehmer (z.B. Name oder Matrikelnummer) zu erfahren. Mehrfachteilnahmen konnten ebenfalls durch die Pseudonymisierungscodes identifiziert und ausgeschlossen werden. Letztlich wurden auch Daten zum Schulerfolg von Teilnehmern mit nicht deutschem Schulabschluss von der Analyse ausgeschlossen, da sowohl Schulausbildung als auch Schulnotencodierung nicht unbedingt vergleichbar sind mit dem deutschen System bzw. Probleme der Übertragbarkeit existierten.

Die Datenauswertung erfolgte mit der SPSS Statistiksoftware (Version 21), wobei die Daten zum I-S-T 2000 R zunächst als Rohdaten von den Antwortbögen übertragen wurden. Unter Verwendung von SPSS-Syntaxen wurden anschließend die vorzunehmenden Recodierungen, die Berechnungen der Skalenwerte sowie die weiteren Analysen durchgeführt.

3 Weitere Instrumente waren enthalten, die zur Klärung anderer Fragestellungen eingesetzte wurden.

4 Der I-S-T 2000 R konnte aus ökonomischen Gründen in Studie 3 leider nicht mehr eingesetzt werden.

# 3 Ergebnisse

Die im ersten Schritt vorgenommene Selektion von sechs Matrizen des HMT für die Kurzversion wurde anhand der Eigenschaften der Items bezüglich Schwierigkeiten, Trennschärfen und den Korrelationen zur I-S-T 2000 R Reasoning-Skala (vgl. Tabelle 2) durchgeführt. Die Items 1 bis 3 konnten übernommen werden, da kein Ausschlusskriterium zur Selektion zutraf. Die Items 4, 6 und 8 wurden dagegen aufgrund der geringen Korrelationen mit dem I-S-T 2000 R ($r$ = .13, $r$ = .05 bzw. $r$ = .08) nicht übernommen. Zudem fiel Item 8 durch eine zu geringe Trennschärfe auf ($r_{it}$ = .22). Zusätzlich zu den bereits selektierten Items 1 bis 3 wurden noch die unproblematischen Items 5, 7 und 9 ausgewählt, wobei die Itemreihenfolge beibehalten wurde. An diesem Punkt wurde die Itemselektion beendet, da a) eine interne Konsistenz von .64 erreicht wurde, b) sowohl einfache und moderat schwierige Aufgaben als auch ein eher schwierigeres Item enthalten waren und zudem c) die mittlere Gesamtbearbeitungsdauer für den Kurztest bei weiteren zusätzlichen Items über zehn Minuten gestiegen wäre.

Die Gesamtdauer des HMT-S lag durchschnittlich bei 9.4 Minuten (vgl. Tabelle 3). Dabei teilt sich die Zeit auf die Instruktion und die eigentliche Aufgabenbearbeitung auf.

In Tabelle 4 sind die Lösungszeiten für die einzelnen Items aufgeführt. Zusätzlich enthält Tabelle 4 die Schwierigkeiten und Trennschärfen der Items. Es zeigen sich weitestgehend Übereinstimmungen der Kennwerte in den drei Studien. Die ersten beiden Items werden relativ schnell gelöst (im Durchschnitt 28 ≤ $M_t$ ≤ 44 Sekunden), Items 3 bis 5 werden im Durchschnitt nach 51 bis 54 Sekunden gelöst und das sechste Item relativ spät (im Durchschnitt 76 ≤ $M_t$ ≤ 77 Sekunden). Zur Schwierigkeit lässt sich feststellen, dass die ersten beiden Items sehr einfach sind (mit Lösungswahrscheinlichkeiten von $p \geq .83$), Items 3 bis 5 weisen mittlere Schwierigkeiten auf ($.53 \leq p \leq .70$) und das letzte Item ist eher schwierig ($p \leq .35$). Die Trennschärfen der Items 1 bis 5 liegen im Bereich von $r_{it}$ = .29 bis $r_{it}$ = .47. Das sechste Item bzw. die letzte Aufgabe fällt von diesem Niveau etwas ab. Die Trennschärfe dieses Items erreicht im günstigsten Fall von Studie 1 einen Wert von $r_{it}$ = .23.

Beim HMT-S werden am häufigsten 4 von 6 Aufgaben richtig gelöst (vgl. Tabelle 5). In allen drei Studien ist die Verteilung der Lösungshäufigkeiten bedingt durch die relative Leichtigkeit linksschief bzw. rechtssteil. Bei der Wölbung zeigt sich ein nicht einheitliches Bild. Während die Verteilungen der Studien 1 und 2 flachgipflig sind, ist die Verteilung in Studie 3 normalgipflig. In Tabelle 5 sind zudem die Indikatoren der internen Konsistenz angeführt, die auf Basis der Kuder-

Tabelle 2 Eigenschaften und I-S-T 2000 R-Korrelationen der HMT-Items in Studie 1

| Item | $M_t$[a] | $p$[b] | $r_{it}$[b] | $r_{IST}$[c] |
|---|---|---|---|---|
| **1** | **43** | **.88** | **.33** | **.21** |
| **2** | **30** | **.85** | **.35** | **.26** |
| **3** | **53** | **.66** | **.40** | **.36** |
| 4 | 57 | .66 | .39 | .13 |
| **5** | **51** | **.65** | **.45** | **.27** |
| 6 | 76 | .54 | .30 | .05 |
| **7** | **53** | **.58** | **.42** | **.33** |
| 8 | 82 | .36 | .22 | .08 |
| **9** | **76** | **.25** | **.35** | **.44** |
| 10 | 68 | .29 | .43 | .35 |
| 11 | 81 | .25 | .29 | .03 |
| 12 | 66 | .30 | .30 | .22 |
| 13 | 66 | .20 | .37 | .31 |
| 14 | 78 | .15 | .24 | .41 |
| 15 | 78 | .16 | .52 | .16 |
| 16 | 60 | .14 | .40 | .03 |
| 17 | 67 | .18 | .28 | .36 |
| 18 | 61 | .15 | .51 | .23 |
| 19 | 52 | .12 | .29 | .13 |
| 20 | 41 | .11 | .44 | .07 |

*I-S-T 2000 R = Intelligenz-Struktur-Test 2000 R (Liepmann et al. 2007). HMT = Hagener Matrizen-Test (Heydasch et al. 2013). $M_t$ = Mittlere Bearbeitungsdauer in Sekunden. p = Schwierigkeit. $r_{it}$ = Trennschärfe. $r_{IST}$ = Itemkorrelation mit dem I-S-T 2000 R Reasoning-Gesamtwert. Fett formatierte Items wurden für die Kurzform selektiert. Unterstrichene Kennwerte führten zum Ausschluss des Items für die Kurzversion ($p \leq .20$, $r_{it} \leq .30$ oder $r_{IST} \leq .20$).*
*[a]$N_1 = 606$. [b]$N_1 = 681$. [c]$N_1 = 56$.*

Tabelle 3 Test- und Bearbeitungsdauer (in Minuten)

| | Percentile | | | | | | |
|---|---|---|---|---|---|---|---|
| | 10% | 25% | 50% | 75% | 90% | $M_t$ | $SD_t$ |
| Instruktion | 1.8 | 2.4 | 3.3 | 4.6 | 6.1 | 4.5 | 9.07 |
| Aufgaben | 2.7 | 3.7 | 4.9 | 3.2 | 7.7 | 5.1 | 1.97 |
| Gesamt | 5.0 | 6.4 | 8.3 | 10.8 | 13.5 | 9.4 | 8.88 |

*Aggregation über alle drei Studien, da keine wesentlichen Unterschiede festzustellen waren. $M_t$ = Mittlere Bearbeitungsdauer. $SD_t$ = Standardabweichung der Bearbeitungsdauer.*
*N = 1.563.*

Tabelle 4 Mittlere Lösungsdauer, Schwierigkeiten und Trennschärfen der Items

| | $M_t$ | | | $p$ | | | $r_{it}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Studie | | | Studie | | | Studie | | |
| Item | 1[a] | 2[b] | 3[c] | 1[d] | 2[e] | 3[f] | 1[d] | 2[e] | 3[f] |
| 1 | 43 | 44 | 39 | .88 | .88 | .88 | .38 | .32 | .34 |
| 2 | 30 | 32 | 28 | .85 | .83 | .87 | .42 | .42 | .35 |
| 3 | 53 | 53 | 52 | .66 | .65 | .67 | .39 | .38 | .25 |
| 4 | 51 | 53 | 51 | .65 | .64 | .70 | .47 | .46 | .39 |
| 5 | 53 | 54 | 53 | .58 | .53 | .57 | .40 | .31 | .29 |
| 6 | 76 | 76 | 77 | .25 | .25 | .35 | .23 | .20 | .21 |

*$M_t$ = Mittlere Bearbeitungsdauer in Sekunden. p = Schwierigkeit. $r_{it}$ = Trennschärfe.*
*[a]$N_1 = 636$. [b]$N_2 = 633$. [c]$N_3 = 229$. [d]$N_1 = 681$. [e]$N_2 = 658$. [f]$N_3 = 233$.*

Tabelle 5 Skaleneigenschaften

| Studie | *N* | *M* | *SD* | Schiefe | Kurtosis | KR20 |
|---|---|---|---|---|---|---|
| 1 | 681 | 3.88 | 1.56 | -0.65 | -0.31 | .64 |
| 2 | 658 | 3.78 | 1.53 | -0.55 | -0.41 | .61 |
| 3 | 233 | 4.03 | 1.47 | -0.76 | 0.05 | .57 |

*KR20 = Interne Konsistenz auf Basis der Kuder-Richardson Formel 20 (Kuder/Richardson 1937).*

Richardson Formel 20 (KR20; Kuder/Richardson 1937) berechnet wurde. Diese schwankt über die Studien hinweg von $KR20_1 = .64$, über $KR20_2 = .61$ bis $KR20_3 = .57$. Wird die interne Konsistenz über alle drei Studien hinweg ermittelt, resultiert ein Wert von KR20 = .62 ($N = 1.572$).

Festzustellen sind Geschlechtsunterschiede (vgl. Tabelle 6). Dieser Effekt ist aber eher klein ($d_1 = 0.16$, $d_2 = 0.25$ und $d_3 = 0.15$; $p_1 = .058$, $p_2 = .004$ und $p_3 = .354$).

Die Beziehung der HMT-S-Testergebnisse zum Alter ist leicht negativ: Ältere Teilnehmer lösen tendenziell weniger Aufgaben. Dieser Zusammenhang ist aber in den ersten beiden Studien niedrig und in Studie 3 nicht nachweisbar ($r_1 = -.13$, $r_2 = -.12$, $r_3 = -.02$; $p_1 = .001$, $p_2 = .003$, $p_3 = .768$; $N_1 = 679$, $N_2 = 654$ bzw. $N_3 = 233$).

Die Korrelationen des HMT-S mit der Langversion betragen $r_1 = .79$ ($p < .001$) in Studie 1 und $r_2 = .78$ in ($p < .001$) Studie 2. Zur inhaltlichen Bestimmung der in der Langversion und der Kurzversion geteilten Varianz wurde für Studie 1 und 2 jeweils eine hierarchische Regression berechnet. Als inhaltlicher Marker und

Tabelle 6 Geschlechtsunterschiede

| Studie | | *N* | *M* | *SD* | *t* | *df* | *d* |
|---|---|---|---|---|---|---|---|
| 1 | Männer | 174 | 4.08 | 1.64 | 1.90 | 676 | 0.16 |
| | Frauen | 504 | 3.82 | 1.51 | | | |
| 2 | Männer | 173 | 4.06 | 1.46 | 2.89** | 316.82 | 0.25 |
| | Frauen | 483 | 3.69 | 1.53 | | | |
| 3 | Männer | 47 | 4.21 | 1.53 | 0.93 | 231 | 0.15 |
| | Frauen | 186 | 3.99 | 1.46 | | | |

*** p < .01.*

Tabelle 7 Hierarchische Regressionen zur inkrementellen Validität des HMT zum HMT-S

| | Studie 1[a] | | | | Studie 2[b] | | | |
|---|---|---|---|---|---|---|---|---|
| | *R* | $\Delta R^2$ | *F* | β | *R* | $\Delta R^2$ | *F* | β |
| Schritt 1 | .530 | .281 | 21.1*** | | .494 | .244 | 10.7** | |
| HMT-S | | | | .530 | | | | .494 |
| Schritt 2 | .567 | .041 | 3.2 | | .607 | .125 | 6.3* | |
| HMT-S | | | | .215 | | | | .072 |
| HMT | | | | .374 | | | | .551* |

*AV: Reasoning-Gesamtwert des Intelligenz-Struktur-Test 2000 R (Liepmann et al. 2007). HMT = Hagener Matrizen-Test (Heydasch et al. 2013). HMT-S = Kurzform des Hagener Matrizen-Test.*
*[a] $N_1$ = 56. [b] $N_2$ = 35. * p < .05. ** p < .01. *** p < .001.*

abhängige Variable diente der Reasoning-Gesamtwert des I-S-T 2000 R. Im ersten Schritt wurde der Testwert des HMT-S als Prädiktor eingegeben und im zweiten Schritt der Testwert des HMT. In Tabelle 7 sind die Ergebnisse der hierarchischen Regressionen festgehalten. Da die Varianzanteile, die über den HMT-S durch den HMT erklärt werden ($\Delta R^2$ = .04 in Studie 1 bzw. $\Delta R^2$ = .13 in Studie 2) im Vergleich zu den Anteilen, die der HMT-S aufklärt ($R^2$ = .28 in Studie 1 bzw. $R^2$ = .24 in Studie 2) relativ klein sind, kann geschlossen werden, dass nicht Fehlervarianz (z.B. Methodenvarianz), sondern die Varianzanteile, die für die Reasoning-Fähigkeit stehen, geteilt werden.

Neben der Validitätsbestimmung durch den Vergleich mit der Langversion wurden Korrelationen zum I-S-T 2000 R berechnet (siehe Tabelle 8). Da sich Kennwerte insbesondere bei den verbalen Intelligenzmaßen von Studie 1 zu Stu-

Tabelle 8 Konvergente Validität zum I-S-T 2000 R

| | | | Studie | | |
|---|---|---|---|---|---|
| Bereich | Inhalt | Aufgaben | 1[a] | 2[b] | 1+2[c] |
| Reasoning | gesamt | | .53*** | .49** | .52*** |
| | $g_f$ | | .50*** | .47** | .49*** |
| | V | | .46*** | .03 | .30** |
| | | SE | .24 | .08 | .16 |
| | | AN | .34* | -.14 | .18 |
| | | GE | .44*** | .13 | .33** |
| | N | | .42** | .53** | .46*** |
| | | RE | .31* | .42* | .35*** |
| | | ZR | .37** | .42* | .39*** |
| | | RZ | .40** | .54*** | .45*** |
| | F | | .47*** | .46** | .47*** |
| | | FA | .25 | .31 | .28** |
| | | WÜ | .50*** | .51** | .50*** |
| | | MA | .21 | .24 | .22* |
| Wissen | gesamt | | .24 | .42* | .31** |
| | $g_c$ | | .15 | .36* | .23* |
| | V | | .06 | .24 | .13 |
| | N | | .26* | .36* | .30** |
| | F | | .29* | .48** | .35*** |
| Merkfähigkeit | gesamt | | .36** | .07 | .24* |
| | V | | .24 | -.04 | .12 |
| | F | | .34** | .16 | .27* |

*I-S-T 2000 R = Intelligenz-Struktur-Test 2000 R (Liepmann et al. 2007); $g_f$ = fluide Intelligenz; V = Verbaler Inhaltsbereich; N = Numerischer Inhaltsbereich; F = Figuraler Inhaltsbereich; $g_c$ = kristalline Intelligenz; SE = Satzergänzungen; AN = Analogien; GE = Gemeinsamkeiten; RE = Rechenaufgaben; ZR = Zahlenreihen; RZ = Rechenzeichen; FA = Figurenauswahl; WÜ = Würfelaufgaben; MA = Matrizen.*
*[a]$N_1 = 56$. [b]$N_2 = 35$. [c]$N_{12} = N_1 + N_2 = 91$. * $p < .05$. ** $p < .01$. *** $p < .001$.*

die 2 deutlich unterschieden, wurden zur besseren Interpretierbarkeit die beiden Stichproben zusammengefasst und die Korrelationen des HMT-S zum I-S-T 2000 R erneut berechnet. So korreliert der HMT-S mit dem Reasoning-Gesamtwert mit $r = .52$ und mit $g_f$ mit $r = .49$. Auf Ebene der verschiedenen Inhaltbereiche liegen die Korrelationen für die Skalen zum numerischen und zum figuralen Reasoning nur knapp unterhalb des Niveaus der Gesamtskala ($r = .46$ bzw. $r = .47$). Der HMT-S

korreliert hingegen mit der verbalen Reasoning-Skala lediglich moderat ($r$ = .30). Die Skalen W und ME korrelieren mit dem HMT-S auf mittlerem Niveau ($r$ = .31 bzw. $r$ = .24). Auf der Ebene der unterschiedlichen Inhaltsbereiche lässt sich ein Zusammenhangsmuster finden, das dem innerhalb des Reasoning-Bereiches vergleichbar ist: die numerischen und die figuralen Skalen korrelieren in etwa gleich hoch auf moderatem Niveau mit dem HMT-S (.27 ≤ $r$ ≤ .35), wohingegen die Skalen mit verbalem Aufgabenmaterial nur schwach mit dem HMT-S korrelieren ($r$ = .13 bei W bzw. $r$ = .12 bei ME). Betrachtet man die Zusammenhänge mit den einzelnen Aufgabengruppen, so wird deutlich, dass die Korrelationen innerhalb eines Inhaltsbereichs heterogen ausfallen. Dies betrifft zwar sowohl die verbalen als auch die numerischen Aufgabengruppen, aber insbesondere die figuralen Aufgaben Figurenauswahl ($r$ = .28), Würfelaufgaben ($r$ = .50) und Matrizen ($r$ = .22) unterscheiden sich hinsichtlich des jeweiligen Zusammenhangs zum HMT-S deutlich.

Die Assoziationen zum ISI (vgl. Tabelle 9) sind über alle drei Studien weitestgehend äquivalent. Positive Zusammenhänge zeigt der HMT-S zu den theoretisch eher nahen Konstrukten mathematische Intelligenz, numerische Intelligenz und logischem Denken. Weitere Korrelationen sind nicht festzustellen, bis auf leicht negative Korrelationen zur körperlich-kinästhetischen, interpersonalen und intrapersonalen Intelligenz. Mittelhohe Korrelationen ergaben sich bei der Prüfung von Zusammenhängen des HMT-S zu Fähigkeitsüberzeugungen im Bereich des mathematischen Selbstkonzepts. Dieser Zusammenhang konnte sowohl in Studie 2 als auch in Studie 3 repliziert werden. Zumindest tendenziell ist ein positiver Zusammenhang zum akademischen Selbstkonzept wie auch eine negative Assoziation zur studiumspezifischen Hilflosigkeit zu identifizieren. Betrachtet man die Korrelationen zu den Dimensionen der AMS-R, so zeigt sich Unabhängigkeit zur Dimension *fear of failure* und allenfalls eine leichte Tendenz eines positiven Zusammenhangs zur Dimension *hope of success*.

Zusammenhänge des HMT-S zu akademischen Erfolgen der Teilnehmer sind in Tabelle 10 aufgeführt. Positive, über alle drei Studien hinweg bestätigte Zusammenhänge bestehen zur Schulabschlussnote sowie zur letzten erhaltenen Note in Mathematik. Uneinheitlich ist die Befundlage bezüglich des Schulabschlussniveaus und der bisher erreichten Durchschnittsnote im Studium. Tendenziell gehen bessere Testwerte mit höherer Schulbildung einher und bessere Testteilnehmer sind erfolgreicher im Studium. Lediglich in jeweils einer der drei Studien resultierten substantielle Ergebnisse bezogen auf das Schulfach Biologie sowie zum Studienfach Methoden und Statistik. Zu den Schulnoten der Fächer Englisch, Deutsch, Kunst und Sport bestehen keine Zusammenhänge.

Tabelle 9 Divergente Validität

| | Studie | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| ISI[a] | | | |
| Verbales Verständnis | .02 | -.07 | -.05 |
| Wortflüssigkeit | -.03 | -.10* | -.04 |
| Mathematische Intelligenz | .23*** | .24*** | .20** |
| Räumliche Intelligenz | .22*** | .20*** | .22*** |
| Gedächtnisfähigkeit | -.04 | -.03 | -.01 |
| Wahrnehmungsgeschwindigkeit | .05 | -.01 | -.07 |
| Logisches Denken | .17*** | .18*** | .14* |
| Musikalische Intelligenz | -.02 | -.04 | -.06 |
| Körperlich-kinästhetisch Intelligenz | -.00 | -.09* | -.13* |
| Interpersonale Intelligenz | -.04 | -.16*** | -.13 |
| Intrapersonale Intelligenz | -.02 | -.14*** | -.05 |
| SWE | | | |
| Allgemein[b] | .09 | .07 | .08 |
| Studiumspezifisch[c] | .18* | .14 | .01 |
| Hilflosigkeit | | | |
| Allgemein[d] | -.14* | .00 | -.02 |
| Studiumspezifisch[c] | -.21** | -.15* | -.10 |
| Selbstkonzept[e] | | | |
| Akademisch | .23*** | .10 | .10 |
| Mathematisch | .28*** | .27*** | .32*** |
| Verbal | .10 | -.07 | -.00 |
| AMS-R[f] | | | |
| *hope of success* | .21*** | .09 | .12 |
| *fear of failure* | -.12 | .04 | -.06 |

*ISI = Inventar zur selbsteingeschätzten Intelligenz (Rammstedt/Rammsayer 2002); SWE = Allgemeine Selbstwirksamkeitserwartung (Schwarzer/Jerusalem 1995); AMS-R = Revised Achievement Motivation Scale (Lang/Fries 2006).* [a]$N_1 = 675$, $N_2 = 654$, $N_3 = 225$. [b]$N_1 = 286$, $N_2 = 262$, $N_3 = 155$. [c]$N_1 = 176$, $N_2 = 177$, $N_3 = 155$. [d]$N_1 = 266$, $N_2 = 242$, $N_3 = 155$. [e]$N_1 = 345$, $N_2 = 317$ $N_3 = 155$. [f]$N_1 = 244$, $N_2 = 242$, $N_3 = 155$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Tabelle 10 Kriteriumsvalidität

| | Studie | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Schulbildung[a] | | | |
| Schulabschlussniveau[b] | .16** | .10 | .00 |
| Abschlussnote | .14* | .16** | .16* |
| Mathematik | .19*** | .20*** | .19** |
| Englisch | .03 | .03 | .09 |
| Deutsch | -.01 | -.02 | .01 |
| Biologie | .07 | .07 | .16* |
| Kunst | .01 | .04 | .09 |
| Sport | .03 | .09 | .03 |
| Studium | | | |
| Durchschnittsnote[c] | .28** | .18* | .06 |
| Note in Methoden und Statistik[d] | .22 | .38** | .05 |

*Noten wurden invertiert, sodass positive Korrelationen bedeuten, dass bessere Testleistungen mit höherer akademischer Leistung assoziiert sind.*
*[a] $N_1 = 308$, $N_2 = 287$, $N_3 = 191$. [b] Allgemeine Hochschulreife = 3, Fachhochschulreife oder fachgebundene Hochschulreife = 2, Mittlerer Schulabschluss = 1 (Spearman-Korrelation). [c] $N_1 = 131$, $N_2 = 124$. $N_3 = 78$. [d] $N_1 = 70$, $N_2 = 70$, $N_3 = 67$. * $p < .05$. ** $p < .01$. *** $p < .001$.*

# 4 Diskussion

Der HMT-S ist ein flexibles Instrument, das bei Online-Erhebungen implementiert (uns sind keine wesentlichen Einschränkungen bei der technischen Umsetzung bekannt) und an beliebigen Orten sowohl im Feld als auch im Labor eingesetzt werden kann, sofern ein internetfähiger PC oder Laptop und Internetzugang zur Verfügung stehen. Dabei ist der HMT-S nicht nur in der psychologischen, sozialwissenschaftlichen oder pädagogischen Forschung einsetzbar, sondern ebenfalls in verschiedenen Praxisfeldern. Die Testanwendung ist nicht auf Psychologinnen und Psychologen beschränkt. Auch anderen nach wissenschaftlichen Standards arbeitenden Personengruppen steht die Nutzung offen. Ausgenommen sind allerdings Anwendungen, bei denen auf Grundlage des HMT-S eine Einzelfalldiagnostik beispielsweise zu Selektionszwecken (z.B. Personalauswahl) erfolgen soll. In diesen Fällen muss auf alternative Verfahren zurückgegriffen werden, die notwendige Gütemaßstäbe erfüllen, wie sie z.B. mit der DIN 33430 (DIN, 2002) definiert sind.

Die Schwierigkeit des HMT-S ist moderat und der Test kann somit Stichproben vorgelegt werden, die weder durch extrem geringe noch durch sehr hohe kognitive Leistungsfähigkeit gekennzeichnet sind. Dennoch kann die Stichprobe durchaus heterogen sein, denn durch die unterschiedlichen Aufgabenschwierigkeiten werden sowohl Boden- als auch Deckeneffekte vermieden: Die Lösungswahrscheinlichkeit nimmt von Aufgabe zu Aufgabe zu und die letzte Aufgabe ist schließlich so schwierig, dass Differenzierungen auf niedrigerem wie auch auf höherem Fähigkeitsniveau möglich sind (im Rahmen eines Instrumentes mit sechs dichotomen Items). Die Ergebnisse zur Schwierigkeit des HMT-S sind allerdings unter Vorbehalt zu interpretieren. Es handelte sich bei den Probanden der vorgestellten Studien um Studierende und folglich ist ein höheres kognitives Fähigkeitsniveau anzunehmen als für die Gesamtpopulation. Dieser Unterschied sollte aber nicht gravierend sein, da die Studierenden nicht wie in anderen Psychologiestudiengängen durch einen Numerus Clausus selektiert wurden.

Der HMT-S ist äußerst ökonomisch. Die Bearbeitungsdauer liegt mit fünf Minuten bei 10% bis 25% bisheriger Matrizen-Tests, die üblicherweise 20 bis 45 Minuten dauern. Die zeitliche Limitierung von zwei Minuten für die Bearbeitung einer einzelnen Aufgabe erscheint dabei in den meisten Fällen ausreichend, um eine Lösungsalternative zu markieren und abzuschicken. Lediglich bei der schwierigeren sechsten und letzten Matrix werden die zwei Minuten Zeitlimit annähernd erreicht. Im Einzelfall kann es aber durchaus der Fall sein, dass die zwölf Minuten Gesamtzeit zur Bearbeitung ausgeschöpft werden[5].

Effekte von Geschlecht und Alter der Teilnehmer auf die Messung sind vorhanden aber gering. Somit deutet sich an, dass der HMT-S bezüglich Geschlecht und Alter ein relativ fairer Test ist. Eine abschließende Bewertung der Befundlage ist aber aufgrund inkonsistenter Befunde zwischen den drei Studien noch nicht möglich.

Die Objektivität sowohl der Auswertung als auch der Interpretation ist gegeben. Die Auswertung erfolgt automatisiert online und die Interpretation erfolgt einheitlich web-basiert und wird den Testteilnehmern als Rückmeldung der individuellen Ergebnisse im Anschluss an die Testaufgaben präsentiert. Die Durchführungsobjektivität ist hingegen nicht unbedingt gegeben (Gnambs/Batinic/Hertel 2011). Die Instruktion, die Darstellung des Tests und die Zeitnahme bei den Aufgaben sind durch die Web-Programmierung objektiv. Wenn jedoch der Test außerhalb einer kontrollierten (Labor-)Umgebung durchgeführt wird, ist die Durch-

5 Dies ist aber nicht zwangsläufig an die Leistungsstärke gebunden. Eine minimale Spearman-Korrelation zwischen dem HMT-S und der Bearbeitungszeit belegt ($r = .08$, $p = .001$, $N = 1.563$), dass – wenn überhaupt – eher bessere Teilnehmer näher am Zeitlimit sind und länger an den Aufgaben arbeiten.

führungsobjektivität gefährdet und es kann zu Störeinflüssen kommen, die z.B. zu einer Unterbrechung der Konzentration oder gar der Testbearbeitung führen. Eine unbeaufsichtigte Feldtestung bietet aber die Möglichkeit, ein hohes Maß an ökologischer Validität zu erreichen, denn die Künstlichkeit einer Laborumgebung mit anwesender Versuchsleitung als Aufsicht ist dann nicht gegeben.

Die Messgenauigkeit des HMT-S ist, nach Maßstäben von Aiken und Groth-Marnat (2006), die für gruppenstatistische Untersuchungen einen Wert von .60 als ausreichend ansehen, mit einer mittleren internen Konsistenz von KR20 = .62 akzeptabel.

Die Validität des HMT-S wird zunächst durch die sehr hohen Korrelationen zur Langversion HMT belegt. Zudem wurde durch die Regressionsanalysen nachgewiesen, dass es sich bei der geteilten Varianz zwischen Lang- und Kurzversion größtenteils um denjenigen Varianzanteil handelt, der die unterschiedlichen Reasoning-Fähigkeiten der Probanden gemessen durch den I-S-T 2000 R erklärt. Damit ist die Übereinstimmung von Kurz- und Langversion weitestgehend auf Intelligenzunterschiede der Probanden zurückzuführen und stellt nicht etwa methodische Artefakte dar, die durch die identische Durchführungsmethodik im unbeaufsichtigten Online-Assessment bei HMT und HMT-S verursacht wurden. Die Regressionsanalysen sprechen zwar für die Überlegenheit der Langversion (immerhin ist diese mit 20 Items mehr als drei Mal so lang), jedoch wird ein Großteil der Varianz des I-S-T 2000 R schon alleine durch den HMT-S aufgeklärt.

Bei den Zusammenhängen des HMT-S mit dem I-S-T 2000 R fallen die Unterschiede in den Koeffizienten zwischen Studie 1 und 2 auf. Hier darf nicht außer Acht gelassen werden, dass die Stichproben eher klein sind ($N_1$ = 56 bzw. $N_2$ = 35) und Schwankungen nicht (ausschließlich) auf die Messungenauigkeit, sondern wohl auf Stichprobenfehler zurückzuführen sind[6]. Die Korrelationen zwischen HMT-S und den verschiedenen Dimensionen des I-S-T 2000 R sind trotz der (auch noch nach der Zusammenfassung) relativ geringen Größe der Stichprobe weitere Belege für die Güte der Kurzversion und bestätigen, dass der HMT-S in der Tat ein Reasoning-Test ist und die Fähigkeit zum schlussfolgernden Denken misst. Inhaltlich bildet die Reasoning-Skala des I-S-T 2000 R im CHC-Modell den Bereich *induction* (I) ab und die Skala numerischen Reasonings ist der Facette *quantitative reasoning* (RQ) zuzuordnen. Die hohen Korrelationen des HMT-S zu diesen Bereichen und geringere Zusammenhänge zu den Bereichen wie Wissen und Merkfähigkeit sind Belege, dass der HMT-S in der Tat eine Operationalisierung von Gf ist, allerdings unter Vernachlässigung deduktiver Fähigkeiten bzw. des *general sequential*

6 Alternative Erklärungen der Unterschiede wie Extremwerte, Fehler bei der Codierung oder bei der Berechnung der Skalenwerte wurden ausgeschlossen.

*reasoning* (RG, siehe Schneider/McGrew 2012). Es lässt sich dabei in den Daten ein Muster erkennen, welches die Bezeichnung des HMT-S als *numerisch-figuralen* Induktions-/Reasoning-Test rechtfertigt: Die Zusammenhänge sind durchgehend höher für numerische und figurale Skalen im Kontrast zu den Skalen mit verbalen Inhalten. Schneider und McGrew (2012) sehen jedoch Gf als nicht inhaltlich gebunden an, genauso wenig lässt sich aus der Beschreibung der Bereiche *induction* (I) und *quantitative reasoning* (RQ) auf figurale Inhalte schließen (auch wenn dies nicht ausgeschlossen wird). Vielmehr ist auf Ebene der 16 breiten Fähigkeitsbereiche der Faktor *visual processing* (Gv) vorgesehen, der die Fähigkeit repräsentiert *„... to make use of simulated mental imagery (often in conjunction with currently perceived images) to solve problems"* (Schneider/McGrew 2012: 129). Die auffällig hohe Korrelation zwischen dem HMT-S und den Würfelaufgaben des I-S-T 2000 R, bei denen Würfel auf Übereinstimmung verglichen werden müssen mit vorgegebenen, gedrehten und gekippten Würfeln, ist vom Aufgabentyp als *visualization* (Vz) zu klassifizieren, dem Kernbereich von Gv (Schneider/McGrew 2012: 129): *„The ability to perceive complex patterns and mentally simulate how they might look when transformed (e.g., rotated, changed in size, partially obscured)"*.

In diesem Zusammenhang ist bemerkenswert, dass die Assoziation zwischen HMT-S und der figuralen Reasoning-Skala im Wesentlichen auf die Würfelaufgaben und nicht auf die Matrizenaufgaben des I-S-T 2000 R zurückzuführen ist. Diese Korrelation fällt mit $r = .22$ relativ gering aus und spricht gegen die Annahme, dass Zusammenhänge auf oberflächlich betrachtet gleiches Aufgabenmaterial – eben figurale Matrizen – zurückzuführen wären. Dies ist möglicherweise auf Unterschiede in den Aufgabenprinzipien zurückzuführen: Im Gegensatz zum HMT-S werden die in den Matrizen enthaltenen mentalen Operationen nicht vollständig genannt oder an Beispielen verdeutlicht. Hinzu kommt, dass insgesamt mehr Regeln bei den I-S-T 2000 R Matrizen zur Anwendung kommen. Nach unserer Auffassung werden damit weitere Intelligenzkomponenten angesprochen (divergentes Denken im Sinne von Einfallsreichtum nach Jäger 1982), die beim HMT-S irrelevant sind und daher die geringe Höhe der Korrelation erklären. Neben der Überprüfung der theoretischen Einordnung einzig durch den I-S-T 2000 R als objektiven Leistungstest sollten zukünftig im Prozess einer fortgeführten Validierung Bezüge zu anderen mehrdimensionalen Testverfahren wie dem Berliner-Intelligenzstruktur-Test (Jäger/Süß/Beauducel 1997) oder dem Wilde-Intelligenz-Test 2 (Kersting/Althoff/Jäger 2008) empirisch geprüft werden.

Auch die Resultate zu den subjektiven Intelligenzselbsteinschätzungen mittels ISI bestätigen die Einordnung des HMT-S als figural-numerischen Reasoning-Test: positive Korrelationen resultierten in den Bereichen mathematische Intelli-

genz, räumliche Intelligenz und logisches Denken. Die Tendenz, dass der HMT-S leicht negativ assoziiert ist mit der körperlich-kinästhetischen, mit der interpersonalen und mit der intrapersonalen Intelligenz könnte auf intraindividuelle Kontrastierung zurückzuführen sein. Erwartungskonform resultierten zudem Zusammenhänge zu den mathematischen und akademischen Bereichen der im Selbstbericht erhobenen Problemlöseerwartungen (SWE), zu denen der Problemlöseerfahrungen (abgebildet in den Selbstkonzepten) und zu denen der Hilflosigkeit. Zur Leistungsmotivation zeichnet sich ein geringer Zusammenhang ab. Diese sind zwar unerwünscht (Stern, 1911), jedoch bei Intelligenztestungen kaum zu vermeiden (vgl. Conrad 1983).

Die Kriteriumsvalidität des HMT-S wird durch die Zusammenhänge zu akademischen Erfolgsmaßen belegt. Es zeichnen sich Zusammenhänge zu Fächern ab, in denen logisches und schlussfolgerndes Denken gefordert ist (im Speziellen sind dies Mathematik bzw. Methoden und Statistik). Geringe Effekte resultieren bezüglich der allgemeineren Maße des akademischen Erfolgs, d.h. für Schulabschlussniveau, Schulabschlussnote und Durchschnittsnote im Studium. Erklärbar ist dies durch die Aggregation von Noten verschiedener Fächer, die unterschiedliche Anforderungen stellen und für die logisch-schlussfolgerndes Denken weniger relevant ist (z.B. für die Schulfächer Sport oder Kunst). Es darf bei den Korrelationen mit den Schulnoten nicht außer Acht gelassen werden, dass a) das zeitliche Intervall zwischen der Testteilnahme am HMT-S und dem Schulabschluss bei den im Durchschnitt über dreißigjährigen Studierenden über zehn Jahre zurück liegt und die Schulnoten nicht in allen Fällen korrekt erinnert werden und b) keine perfekte Stabilität des erfassten Konstrukts vorliegt. Da die Stichproben der drei berichteten Studien nicht repräsentativ und relativ homogen sind, sollten zur weiteren Validierung Untersuchungen z.B. mit Schülerinnen und Schülern stattfinden, deren aktuelle Noten ggf. objektiv und nicht im Selbstbericht erfasst werden könnten. Auch Studierende anderer Fachrichtungen könnten zukünftige Stichproben bilden, wobei insgesamt eine ausgewogenere Geschlechterverteilung anzustreben wäre.

Unser Fazit zum HMT-S fällt auch unter Berücksichtigung der Einschränkungen der Studien aufgrund relativ homogener und bezüglich der Zusammenhangsmaße mit dem I-S-T 2000 R eher kleinen Stichproben insgesamt positiv aus. Der HMT-S ist ein höchst ökonomischer, kostenloser Online-Intelligenztest, der figural-numerisches Reasoning erfasst. Da nicht nur die Korrelation zur Originalversion sehr hoch, sondern auch das Muster der Validitätskoeffizienten mit dem des HMT vergleichbar ist, kann davon ausgegangen werden, dass der HMT-S den HMT auch auf inhaltlicher Ebene abbildet. Die Reliabilität des HMT-S ist trotz der niedrigen Anzahl von insgesamt nur sechs Aufgaben zufriedenstellend und für Grup-

penuntersuchungen geeignet. Die in den drei Studien gesammelten Belege für die Konstrukt- und die Kriteriumsvalidität sind vielversprechend und wir erhoffen uns zukünftig eine weite Verbreitung und Etablierung des HMT-S. Angefordert werden kann der HMT-S über die URL http://HMT.de.lv[7], über die auch weitere Informationen zugänglich sind.

## Literatur

Aiken, L. R. und G. Groth-Marnat, 2006: Psychological Testing and Assessment (12. Auflage). Boston, MA: Pearson Education.

Amthauer, R., B. Brocke, D. Liepmann und A. Beauducel, 2001: I-S-T 2000 R. Intelligenz-Struktur-Test 2000 R. Göttingen: Hogrefe.

Blickle, G., J. Kramer und J. Mierke, 2010: Telephone-Administered Intelligence Testing for Research in Work and Organizational Psychology. A Comparative Assessment Study. European Journal of Psychological Assessment 26: 154-161.

Buchwald, F., C. Spoden, J. Fleischer und D. Leutner, 2013: Verzweigte Lernumgebungen und Tests mit EFS Survey 8. Diagnostica 59: 113-117.

Conrad, W., 1983: Intelligenzdiagnostik. S. 104-201 in: K.-J. Groffmann und L. Michel (Hg.): Intelligenz- und Leistungsdiagnostik, Enzyklopädie der Psychologie (Themenbereich B, Serie II, Band 2). Göttingen: Hogrefe.

Carroll, J. B., 1993: Human Cognitive Abilities: A Survey of Factoranalytic Studies. Cambridge, NY: University Press.

Carroll, J. B., 2005: The Three-Stratum Theory of Cognitive Abilities. S. 69-76 in: D. P. Flanagan und L. Harrison (Hg.): Contemporary Intellectual Assessment: Theories, Tests, and Issues (2. Auflage). New York, NY: Guilford Press.

Cattell, R. B., 1957: Personality and Motivation Structure and Measurement. New York, NY: World Book.

Cattell, R. B., 1987: Intelligence: Its Structure, Growth and Action. Amsterdam: Elsevier.

Davidson, J. E. und I. A. Kemp, 2011: Contemporary Models. S. 58-82 in: R. J. Sternberg und S. B. Kaufman (Hg.): The Cambridge Handbook of Intelligence. Cambridge, NY: University Press.

Deutsche Gesellschaft für Psychologie & Bundesverband Deutscher Psychologinnen und Psychologen, 2004: Ethische Richtlinien der DGPs und des BDP. http://www.dgps.de/dgps/aufgaben/003.php

DIN, 2002: DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen. Berlin: Beuth.

Evans, J. St. B. T., 2005: Deductive Reasoning. S. 169-184 in: K. J. Holyoak und R. G. Morrison (Hg.): The Cambridge Handbook of Thinking and Reasoning. Cambridge, NY: University Press.

Gardner, G., 1983: Frames of Mind: The Theory of Multiple Intelligences. New York, NY: Basic Books.

Gnambs, T., B. Batinic und G. Hertel, 2011: Internetbasierte psychologische Diagnostik. S. 448-498 in: L. F. Hornke, M. Amelang und M. Kersting (Hg.): Verfahren zur Leistungs-,

7 Bei der URL handelt es sich um eine Weiterleitung zu http://www.fernuni-hagen.de/psychologie/mde/forschung/hot_project.shtml

Intelligenz- und Verhaltensdiagnostik, Enzyklopädie der Psychologie, Psychologische Diagnostik (Bd. II/3). Göttingen: Hogrefe.
Guilford, J. P., 1967: The Nature of Intelligence. New York, NY: McGraw-Hill.
Guttman, L., 1965: A Faceted Definition of Intelligence. S. 166-181 in: R. Eiferman (Hg.): Studies in Psychology, scripta hierosolymitana (Vol. 14). Jerusalem: Hebrew University.
Guttman, L. und S. Levy, 1991: Two Structural Laws for Intelligence Tests. Intelligence 15: 79-103.
Gottfredson, L. S., 1997: Mainstream Science on Intelligence: An Editorial with 52 Signatories, History, and Bibliography. Intelligence 24: 13-23.
Gottfredson, L. S. und D. H. Saklofske, 2009: Intelligence: Foundations and Issues in Assessment. Canadian Psychology 50: 183-195.
Harrell, T. W. und M. S. Harrell, 1945: Army General Classification Test Scores for Civilian Occupations. Educational and Psychological Measurement 5: 229-239.
Heydasch, T., K.-H. Renner, J. Haubrich, B. Hilbig und I. Zettler, 2013: A Free Web-Based Intelligence Test: The Hagen Matrices Test (HMT). Manuskript in Vorbereitung.
Horn, J. L. und A. N. Blankson, 2012: Foundations for Better Understanding of Cognitive Abilities. S. 73-98 in: D. Flanagan und P. Harrison (Hg.): Contemporary Intellectual Assessment: Theories, Tests, and Issues (3. Auflage). New York, NY: Guilford Press.
Horn, J. L. und J. Noll, 1997: Human Cognitive Capabilities: Gf-Gc Theory. S. 53-91 in: D. P. Flanagan, J. L. Gensaft und P. L. Harrison (Hg.): Contemporary Intellectual Assessment: Theories, Tests, and Issues. New York, NY: Guilford Press.
Hülsheger, U. R., G. W. Maier und T. Stumpp, 2007: Validity of General Mental Ability for the Prediction of Job Performance and Training Success in Germany: A Meta-Analysis. International Journal of Selection and Assessment 15: 3-18.
Hunter, J. E. und R. F. Hunter, 1984: Validity and Utility of Alternative Predictors of Job Performance. Psychological Bulletin 96: 72-98.
Jäger, A. O., 1982: Mehrmodale Klassifikation von Intelligenzleistungen: Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells. Diagnostica 28: 195-225.
Jäger, A. O., H.-M. Süß und A. Beauducel, 1997: Berliner Intelligenzstruktur - Test. Form 4. Göttingen: Hogrefe.
Jensen, A., 1998: The g Factor: The Science of Mental Ability. Westport, CT: Praeger.
Jerusalem, M. und R. Schwarzer, 1986: Fragebogen zur Erfassung von Hilflosigkeit. In: R. Schwarzer (Hg.): Skalen zur Befindlichkeit und Persönlichkeit (Forschungsbericht 5). Berlin FU Berlin, Institut für Psychologie, Abt. Pädagogische Psychologie.
Jerusalem, M. und R. Schwarzer, 2012: Dimensionen der Hilflosigkeit. In: A. Glöckner-Rist (Hg.): Zusammenstellung sozialwissenschaftlicher Items und Skalen. ZIS Version 15.00. Bonn: GESIS.
Judge, T. A., C. A. Higgins, C. J. Thoresen und M. R. Barrick, 1999: The Big Five Personality Traits, General Mental Ability, and Career Success Across the Life Span. Personnel Psychology 52: 621-652.
Kersting, M., K. Althoff und A. O. Jäger, 2008: WIT-2. Der Wilde-Intelligenztest. Göttingen: Hogrefe.
Khandaker, G. M., J. H. Barnett, I. R. White und P. B. Jones, 2011: A Quantitative Meta-Analysis of Population-Based Studies of Premorbid Intelligence and Schizophrenia. Schizophrenia Research 132: 220-227.
Kuder, G. F. und M. W. Richardson, 1937: The Theory of the Estimation of Test Reliability. Psychometrika 2: 151-160.

Kuncel, N. R., S. A. Hezlett und D. S. Ones, 2004: Academic Performance, Career Potential, Creativity, and Job Performance: Can One Construct Predict Them All? Journal of Personality and Social Psychology 86: 148-161.

Kunina, O., O. Wilhelm, M. Formazin, K. Jonkmann und U. Schroeders, 2007: Extended Criteria and Predictors in College Admission: Exploring the Structure of Study Success and Investigating the Validity of Domain Knowledge. Psychology Science 49: 88-114.

Lang, J. W. B. und S. Fries, 2006: A Revised 10-Item Version of the Achievement Motives Scale. Psychometric Properties in German-Speaking Samples. European Journal of Psychological Assessment 22: 216-224.

Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID) (Hg.), 2012: Verzeichnis Testverfahren. Kurznamen. Langnamen. Autoren. Testrezensionen (19., aktualisierte Auflage). Trier: ZPID.

Liepmann, D., A. Beauducel, B. Brocke und R. Amthauer, 2007: Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R). Manual (2. Auflage). Göttingen: Hogrefe.

McGrew, K. S., 1997: Analysis of the Major Intelligence Batteries According to a Proposed Comprehensive Gf-Gc Framework. S. 151-179 in: D. P. Flanagan, J. L. Genshaft und P. L. Harrison (Hg.): Contemporary Intellectual Assessment: Theories, Tests, and Issues. New York: Guilford Press.

McGrew, K. S., 2005: The Cattell Horn Carroll (CHC) Theory of Cognitive Abilities. Past, Present and Future. S. 136-202 in: D. P. Flanagan und P. L. Harrison (Hg.): Contemporary Intellectual Assessment: Theories, Tests, and Issues (2. Auflage). New York, NY: Guilford Press.

Ng, T. W. H., L. T. Eby, K. L. Sorensen und D. C. Feldman, 2005: Predictors of Objective and Subjective Career Success: A Meta-Analysis. Personnel Psychology 58: 367-408.

Pawlik, K., 2006: Psychologische Diagnostik I: Methodische Grundlagen. S. 555-562 in: K. Pawlik (Hg.): Handbuch Psychologie. Wissenschaft–Anwendung-Berufsfelder. Berlin: Springer.

Poropat, A. E., 2009: A Meta-Analysis of the Five-Factor Model of Personality and Academic Performance. Psychological Bulletin 135: 322-338.

Rammstedt, B. und T. Rammsayer, 2002: Die Erfassung der selbsteingeschätzten Intelligenz: Konstruktion, teststatistische Überprüfung und erste Ergebnisse des Inventars zur selbsteingeschätzten Intelligenz (ISI). Zeitschrift für Differentielle und Diagnostische Psychologie 23: 435-466.

Rindermann, H. und A. Neubauer, 2000: Informationsverarbeitungsgeschwindigkeit und Schulerfolg: Weisen basale Maße der Intelligenz prädiktive Validität auf? Diagnostica 46: 8-17.

Roberts, B. W., N. R. Kuncel, R. Shiner, A. Caspi und L.R. Goldberg, 2007: The Power of Personality: The Comparative Validity of Personality Traits, Socioeconomic Status, and Cognitive Ability for Predicting Important Life Outcomes. Perspectives on Psychological Science 2: 313-345.

Salgado, J. F., N. Anderson, S. Moscoso, C. Bertua und F. de Fruyt, 2003: International Validity Generalization of GMA and Cognitive Abilities: A European Community Meta-Analysis. Personnel Psychology 56: 573-605.

Schiefele, U., B. Moschner und R. Husstegge, 2002: Skalenhandbuch SMILE-Projekt (unveröffentlichtes Manuskript). Bielefeld: University.

Schmidt, F. L. und J. E. Hunter, 1998: The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings. Psychological Bulletin 124: 262-274.

Schmidt-Atzert, L., 2002. Intelligenz-Struktur-Test 2000 R (Testrezension). Zeitschrift für Personalpsychologie 1: 50-56.

Schmidt-Atzert, L. und W. Rauch, 2008: Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R). 2., erweiterte und überarbeitete Auflage [TBS-TK Rezension]. Report Psychologie 33: 303-304.

Schneider, W. J. und K. S. McGrew, 2012: The Cattell-Horn-Carroll Model of Intelligence. S. 99-144 in: D. Flanagan und P. Harrison (Hg.): Contemporary Intellectual Assessment: Theories, Tests, and Issues (3. Auflage). New York, NY: Guilford Press.

Schwarzer, R. und M. Jerusalem, 1995: Generalized Self-Efficacy Scale. S. 35-37 in: J. Weinman, S. Wright und M. Johnston (Hg.): Measures in Health Psychology: A User's Portfolio. Causal and Control Beliefs. Windsor: NFER-Nelson.

Seligman, M. E. P., 1975: Helplessness: On Depression, Development and Death. San Francisco, CA: Freeman.

Silverstein, A. B., 1990: Short Forms of Individual Intelligence Tests. Psychological Assessment: A Journal of Consulting and Clinical Psychology 2: 3-11.

Sloman, S. A. und D. A. Lagnado, 2005: The Problem of Induction. S. 95-116 in: K. J. Holyoak und R. G. Morrison (Hg.): The Cambridge Handbook of Thinking and Reasoning. Cambridge, NY: University Press.

Spearman, C., 1904: The Proof and Measurement of Association between Two Things. The American Journal of Psychology 15: 72-101.

Stern, W., 1911: Die Differentielle Psychologie in ihren methodischen Grundlagen. Leipzig: Johann Ambrosius Barth.

Sternberg, R. J., 2005: Intelligence. S. 751-773 in: K. J. Holyoak und R. G. Morrison (Hg.): The Cambridge Handbook of Thinking and Reasoning. Cambridge, NY: University Press.

Strenze, T., 2007: Intelligence and Socioeconomic Success: A Meta-Analytic Review of Longitudinal Research. Intelligence 35: 401-426.

Süß, H.-M. und A. Beauducel, 2004: Faceted Models of Intelligence. S. 313-332 in: O. Wilhelm und R. W. Engle (Hg.): Handbook of Understanding and Measuring Intelligence. Thousand Oaks, CA: Sage.

Wasserman, J. D., 2012: A History of Intelligence Assessment: The Unfinished Tapestry. S. 3-55 in: D. P. Flanagan und P. L. Harrison (Hg.): Contemporary Intellectual Assessment: Theories, Tests, and Issues (3. Auflage). Cambridge, NY: University Press.

Willis, J. O., R. Dumont und A. S. Kaufman, 2011: Factor-Analytic Models of Intelligence. S. 39-57 in: R. J. Sternberg und S. B. Kaufman (Hg.): The Cambridge Handbook of Intelligence. Cambridge, NY: University Press.

Ziegler, M., E. Dietl, E. Danay, M. Vogel und M. Bühner, 2011: Predicting Training Success with General Mental Ability, Specific Ability Tests, and (Un)Structured Interviews: A meta-analysis with unique samples. International Journal of Selection and Assessment 19: 170-182.

| | |
|---|---|
| Anschrift der Autoren | Timo Heydasch<br>FernUniversität in Hagen<br>Universitätsstr. 33<br>58084 Hagen<br>Timo.Heydasch@FernUni-Hagen.de |
| Ko-Autor/-in | Julia Haubrich<br>FernUniversität in Hagen |
| | Karl-Heinz Renner<br>FernUniversität in Hagen |

methoden, daten, analysen · 2013, Jg. 7(2), S. 209-232 DOI: 10.12758/mda.2013.012

# Konstruktion und Validierung einer Skala zur relativen Messung von physischer Attraktivität mit einem Item

## *Das Attraktivitätsrating 1 (AR1)*

# Development and validation of a single-item scale for the relative assessment of physical attractiveness

## *The Attractiveness Rating 1 (AR1)*

*Johannes Lutz, Christoph J. Kemper, Constanze Beierlein, Jutta Margraf-Stiksrud und Beatrice Rammstedt*

*Zusammenfassung*

Physische Attraktivität ist ein wichtiger Faktor sozialer Interaktion und Kognition, z.B. bei der Partnerwahl oder dem Eindruck, den sich Menschen von sich selbst und von anderen bilden. Auch im Rahmen von Umfragen könnte die physische Attraktivität sozial- und gesundheitswissenschaftliche Erklärungsmodelle sinnvoll ergänzen. So finden sich unter anderem Zusammenhänge mit Kriterien des Berufserfolgs, diversen Gesundheitsvariablen und der allgemeinen Lebenszufriedenheit. Die Messung physischer Attraktivität mittels absoluten Einschätzungen hat sich jedoch als problematisch erwiesen. Attraktivitätsurteile können z.B. von Charakteristika der Beurteiler beeinflusst werden. Vor allem die Interaktion zwischen Alter und Geschlecht von Beurteiler und Zielperson wird in der Literatur als Einflussfaktor berichtet. Ziel

*Abstract*

Physical attractiveness is an important factor of social interaction and cognition, such as in mate choice or in the impressions people form of themselves and of others. In the context of social and health sciences, physical attractiveness has also been shown to complement the prediction of various target variables. Among other relationships, physical attractiveness is associated with criteria of professional success, various health variables and overall life satisfaction. However, the measurement of physical attractiveness via absolute ratings has several shortcomings. For example, attractiveness judgments are influenced by characteristics of the rater. As pointed out in literature, especially the interaction between age and gender of rater and target person has a significant impact on rating scores. The aim of this study was to construct a scale for the

der vorliegenden Studie war daher die Konstruktion einer Skala zur ökonomischen, reliablen und validen Messung von physischer Attraktivität, die den in der Literatur berichteten Problemen Rechnung trägt. Um dies zu erreichen wird im Attraktivitätsrating 1 (AR1) ein relatives Messkonzept eingesetzt. Das AR1 wurde anhand von vier alters-, geschlechts- und bildungsheterogenen Stichproben konstruiert und validiert. Die berichteten Ergebnisse sprechen dafür, dass das AR1 das Merkmal physische Attraktivität adäquat abbildet.

economical, reliable and valid assessment of physical attractiveness, controlling for methodological problems reported in prior research. To achieve this, the Attractiveness Rating 1 (AR1) employs a relative measurement approach. The AR1 was constructed and validated using four diverse samples. The presented results suggest that the AR1 adequately measures physical attractiveness.

## 1 Theoretischer Hintergrund

Physische Attraktivität ist ein wichtiges Personenmerkmal zur Beschreibung und Erklärung menschlichen Erlebens und Verhaltens. Unter physischer Attraktivität wird die Attraktivität einer Person verstanden, die auf ihren körperlichen Merkmalen beruht (Asendorpf 2011, S.63). Die physische Attraktivität einer Person wird durch den Betrachter aus Merkmalen des Gesichts, wie der Größe der Nase, der Augen und des Kinns (Kindchenschema, vgl. Lorenz 1943) und aus der Figur, z.B. dem Taille-Hüfte-Verhältnis, erschlossen. Sie hat weitreichende Implikationen für die zwischenmenschliche Interaktion. Die physische Attraktivität ist wesentlicher Bestandteil des Selbstkonzepts und beeinflusst, wie sich eine Person selbst erlebt und somit auch wie sie sich anderen gegenüber verhält. Darüber hinaus beeinflusst die physische Attraktivität auch wie eine Person von ihren Interaktionspartnern gesehen und behandelt wird. Physisch attraktivere Menschen werden weniger attraktiven oder gar körperlich fehlgebildeten Menschen gegenüber positiver eingeschätzt und in der sozialen Interaktion, z.B. bei der Partnerwahl oder bei Bewerbungen, bevorzugt. Bestrebungen, dem gesellschaftlichen Schönheitsideal durch den Kauf modischer Kleidung und kosmetischer Produkte, durch Diäten, Sport, Sonnenbaden, Bodybuilding usw. näher zu kommen, erscheinen aufgrund der vorteilhaften Wirkung eines attraktiven Äußeren nur allzu verständlich. Aufgrund ihrer Rolle für das menschliche Erleben und Verhalten bei einer Vielzahl sozialer Phänomene und Prozesse ist die physische Attraktivität für die sozialwissenschaftliche Forschung und darüber hinaus ein hochgradig relevantes Konstrukt. Ein ökonomisches und reliables Verfahren für die Messung dieses Konstrukts zu entwickeln und der Forschung zur Verfügung zu stellen, ist Ziel der hier vorgestellten Arbeit.

Der Eindruck physischer Attraktivität entsteht in der sozialen Interaktion. Das Wahrnehmen von Proportionen, Haarfarbe, Frisur und mimischen Ausdrucks des Gesichts, löst auf Seiten des Betrachters einer Person zahlreiche Einschätzungsvorgänge aus, die in einer Zuschreibung von physischer Attraktivität resultieren. Dabei spielen insbesondere Gesichts- und Figurmerkmale eine Rolle (Gallup/Frederick 2010). Die Wahrnehmung physischer Attraktivität wird bspw. durch die Körperkomposition und die Beschaffenheit der Haut beeinflusst. So werden Frauen mit niedrigem Taille-Hüfte-Verhältnis und besonders ebenmäßiger Haut als attraktiver wahrgenommen (Fink et al. 2001; Singh et al. 2010). Bei Männern wirkt ein höheres Schulter-Hüfte-Verhältnis und moderat ausgeprägte Muskulatur attraktiv (Dijkstra/Buunk 2001; Frederick/Haselton 2007). Neben den genannten Merkmalen kommt dem Gesicht eine besondere Bedeutung für die physische Attraktivität zu. Dies wird bereits durch das heute nur noch selten verwendete Synonym „Antlitz" impliziert. Es leitet sich aus „entgegenblicken" ab. „Ein Gesicht wirkt dadurch, wie es dem Betrachter entgegenblickt" (Margraf-Stiksrud 1991, S. 11). Die beiden Aspekte des Gesichts, die Mimik als dynamische Qualität und die Physiognomie als statische Qualität, sind wichtige Informationsquellen in der sozialen Interaktion (Margraf-Stiksrud 1991; Sergl 1991). So dient der mimische Ausdruck primär der Aussendung von Signalen über Stimmungen und innere Zustände und somit der zwischenmenschlichen Kommunikation. Die Physiognomie vermittelt Interaktionspartnern ebenfalls Informationen, auch z.B. über die Fruchtbarkeit einer Person (vgl. Rhodes 2006). Die Ausbildung bestimmter physiognomischer Merkmale, z.B. markantes Kinn, vorspringende Backenknochen, Gesichtsbehaarung bei Männern und volle Lippen, hohe Wangenknochen, schmale Wangen bei Frauen, wird durch Geschlechtshormone gesteuert, die auch als Indikatoren für ein starkes Immunsystem gelten (vgl. Rhodes 2006). Aus evolutionsbiologischer Perspektive lässt sich argumentieren, dass diese Merkmale biologisch verankerte Signale für Fruchtbarkeit darstellen, deren Erkennen daher einen Selektionsvorteil darstellt. Man vermutet, dass sie aus diesem Grund auch als attraktiv wahrgenommen werden (Bierhoff 2000; Rhodes 2006). Neben den oben genannten Merkmalen, die sich bei Männern und Frauen im Hinblick auf die Zuschreibung von physischer Attraktivität unterscheiden (Geschlechtsdimorphismus), gelten bei beiden Geschlechtern Symmetrie der Gesichtshälften und Durchschnittlichkeit der physiognomischen Merkmale als relevante Prädiktoren für physische Attraktivität (Rhodes 2006).

Die Auswirkungen der physischen Attraktivität auf Betrachter und Betrachtete sind in vielen Bereichen substantiell. Viele interpersonelle Prozesse werden durch physische Attraktivität beeinflusst. Zahlreiche Studien belegen, dass Attraktive im Vergleich zu weniger Attraktiven in der sozialen Interaktion oft bevorzugt

werden und Vorteile haben: Attraktivere bekommen häufiger als weniger Attraktive gut bezahlte berufliche Positionen mit hohem Prestige (z.B. Schuler/Berger 1979; Umberson/Hughes 1987). Attraktivere Kinder werden seltener von Lehrern diszipliniert (Clifford/Walster 1973). Attraktivere Kriminelle erhalten mildere Urteile als weniger Attraktive (Sigall/Ostrove 1975). Attraktiveren wird mehr Aufmerksamkeit geschenkt und ihnen wird häufiger geholfen (Langlois et al. 2000). Attraktivere haben weiterhin diverse Vorteile in der sozialen Interaktion, die dem Aufbau intimer Beziehungen zuträglich sind (Tramitz 2000). Solche Befunde lassen sich zum Teil auf ein Attraktivitätsstereotyp zurückführen. Menschen neigen dazu, physisch Attraktiven sozial erwünschte Eigenschaften wie „interessant", „warm", „unabhängig" oder „ehrlich" zuzuschreiben und außerdem mehr Lebensglück, beruflichen Erfolg und soziale Kompetenz (Bierhoff 2000; Eagly et al. 1991). Neben diesen Zuschreibungen, gemäß der Heuristik „schön ist gut", weisen Attraktive aber tatsächlich viele sozial erwünschte Eigenschaften auf. Sie sind z.B. geselliger und beliebter (Langlois et al. 2000). Außerdem steht physische Attraktivität in positivem Zusammenhang mit der allgemeinen Gesundheit (vgl. Thornhill/Gangestad 1999), der Anzahl der Nachkommen (Jokela 2009), der Intelligenz (Langlois et al. 2000; Kanazawa/Kovar 2004), dem durchschnittlichen Einkommen (Frieze et al. 1991; Judge et al. 2009) und verschiedenen anderen Kriterien des Berufserfolgs (Hamermesh/Biddle 1994; Hosoda et al. 2003).

Während Attraktive häufiger positive Reaktionen aus ihrem sozialen Umfeld erfahren und Vorteile in der sozialen Interaktion genießen, erfahren Menschen mit geringer Attraktivität zum Teil negative Reaktionen. Dies trifft besonders für Menschen zu, die stark von der Idealnorm abweichen, die z.B. Fehlbildungen im Gesicht aufweisen. Menschen mit normabweichenden physiognomischen Merkmalen werden oft mit Vorurteilen konfrontiert und in der sozialen Interaktion benachteiligt (siehe Heimes/Kemper 2001): Zu Menschen mit entstellenden Gesichtsmerkmalen wird eine größere soziale Distanz gehalten, z.B. beim Warten an der Bushaltestelle (Fleischer-Peters/Margraf-Stiksrud 1996). Houston und Bull (1994) beobachteten, dass Plätze in der Straßenbahn neben Menschen mit einem Feuermal (Nevus flammeus) im Gesicht häufiger gemieden werden. Neben einer größeren sozialen Distanz wurden auch Vorurteile gegenüber Personen mit Fehlbildungen berichtet, z.B. bei Personen mit schlecht operierter Lippen-Kiefer-Gaumenspalte (Bernstein 1982; Sergl 1991). Für die Vorurteile und das ablehnende Verhalten lassen sich verschiedene Gründe anführen. Unter anderem gehen die Normabweichungen in der Physiognomie mit einer Mimik einher, die dem Beobachter das Erkennen ausgesendeter Signale erschwert. Daraus resultiert eine Kommunikation, die von einem Inter-

aktionspartner nicht mehr eindeutig interpretiert werden kann (Fleischer-Peters/Margraf-Stiksrud 1996).

Die berichteten Befunde zeigen, dass dem Gesicht eine bedeutende Rolle bei der zwischenmenschlichen Begegnung zukommt. Dies scheint im Besonderen für das attraktive und „normkonforme" Gesicht zu gelten, da es wichtige Voraussetzungen für eine reibungslose Kommunikation schafft. Das Gesicht daher bei der Messung des Konstrukts physische Attraktivität ins Zentrum zu stellen, liegt demnach nahe. Die Praxis der Attraktivitätsmessung sieht aktuell allerdings anders aus. Üblicherweise wird physische Attraktivität mittels mehrstufiger Ratingskalen mit verbalen Ankern, z.B. von „gar nicht attraktiv" bis „sehr attraktiv" gemessen (Langlois et al. 2000). Dabei wird von den Beuteilern eine globale Einschätzung gefordert, d.h. es wird nicht spezifiziert, welcher Aspekt von Attraktivität beurteilt werden soll. Welche äußerlichen Merkmale der Beurteiler letzten Endes berücksichtigt, bleibt dabei unklar. Diese Art der Messung, die z.B. im ALLBUS (Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften) eingesetzt wird, weist daher nicht nur eine Unterbetonung der Bedeutung des Gesichts auf, sondern auch eine mangelnde Standardisierung der Attraktivitätseinschätzung.

Weiterhin gehen mit der Messung von physischer Attraktivität Herausforderungen einher, die durch diese Art der Messung nicht adäquat adressiert werden. Attraktivitätseinschätzungen hängen von verschiedenen Merkmalen des Beurteilers ab, z.B. von seinem Vergleichsstandard, seinem Geschlecht und seinem Alter (Bierhoff 2000; Henss 1991; Hönekopp 2006; Nedelec/Beaver 2011). Männer, die kurz zuvor Filme mit attraktiven Frauen oder Aktfotographien angesehen haben, beurteilten durchschnittlich attraktive Frauen und ihre eigenen Frauen als weniger attraktiv im Vergleich zu Männern, bei denen der Vergleichsmaßstab nicht experimentell manipuliert wurde (Kontrasteffekt; Kenrick/Gutierres 1980; Kenrick et al. 1989). Nedelec und Beaver (2011) berichten geschlechtsabhängige Unterschiede in der Attraktivitätseinschätzung von Männern. Während Männer bei der Einschätzung von Männern eher zu mittleren Urteilen neigen, beurteilen Frauen die Männer häufiger entweder als „sehr attraktiv" oder „sehr unattraktiv". Der generell negative Zusammenhang zwischen Alter und physischer Attraktivität wird ebenfalls durch das Geschlecht und auch das Alter des Beurteilers beeinflusst (Henss 1991). Bei der Beurteilung von Personen des eigenen Geschlechts oder der eigenen Altersgruppe fällt der Zusammenhang zwischen Alter und physischer Attraktivität signifikant niedriger aus. Der Effekt, dass Gesichter von Personen der eigenen Altersgruppe besser wiedererkannt (und somit auch anders eingeschätzt) werden als Gesichter von älteren oder jüngeren Menschen, wird als „Own-Age-Bias" (OAB) bezeichnet (vgl. Bartlett/Fulton 1991; Rhodes/Anastasi 2012).

Diese Befunde verdeutlichen, welche Verzerrungen bei der üblichen Messung ohne Vergleichsmaßstab (fortan: „absolute" Attraktivitätseinschätzung) auftreten können. Ziel der hier vorgestellten Arbeit war es, eine alternative Ratingskala (Attraktivitätsrating 1, AR1) zu konstruieren, die den oben beschriebenen Verzerrungen Rechnung trägt und diese minimiert. Zudem sollte diese optimal auf die Erfordernisse der sozialwissenschaftlichen Forschung ausgerichtet sein, d.h. eine möglichst ökonomische aber dennoch ausreichend reliable und valide Messung des Konstrukts erlauben. Dazu wird im AR1 ein „relatives Messkonzept" verwendet, bei dem die Einschätzung der Attraktivität einer Person (fortan auch Zielperson) an einem Vergleichsmaßstab erfolgt. Als Vergleichsmaßstab dient ein „Ankerbild" mit dem Gesicht einer durchschnittlich attraktiven Person (ungefähr) desselben Alters und desselben Geschlechts wie die Zielperson. Die Attraktivität der Zielperson soll vom Beurteiler relativ zum Ankerbild eingeschätzt werden. Durch dieses relative Messkonzept wird (1) die Grundlage von Attraktivitätseinschätzungen vereinheitlicht und die Vergleichbarkeit über Beurteiler erhöht, (2) der besonderen Rolle des Gesichts Rechnung getragen und (3) Verzerrungen der Attraktivitätsmessung durch Merkmale der Beurteiler, wie deren individuelle Präferenzen, deren Alter und deren Geschlecht, minimiert. Im Folgenden werden Konstruktion und empirische Überprüfung des AR1 dargestellt.

# 2 Methode

## 2.1 Stichproben

Das AR1 wurde anhand von vier alters-, geschlechts- und bildungsheterogenen Stichproben entwickelt und validiert. Details zur Zusammensetzung der Stichproben sind in Tabelle 1 dargestellt. Die Erhebung der Stichproben 1, 2 und 3 wurden von unabhängigen kommerziellen Anbietern durchgeführt.

Stichprobe 1 ist eine Quotenstichprobe, geschichtet nach den Merkmalen Geschlecht, Alter, Bildung und Bundesland ($N$ = 539). Die Grundgesamtheit war definiert als „alle in der Bundesrepublik Deutschland in Privathaushalten lebenden deutschsprachigen Personen ab 18 Jahren." Die Daten wurden im Rahmen eines persönlich-mündlichen Interviews (CAPI, Computer Assisted Personal Interview) und durch die Vorgabe eines Papierfragebogens erhoben. Die Erhebung dauerte insgesamt im Mittel 53 Minuten ($SD$ = 12).

Stichprobe 2 wurde im Internet erhoben (CAWI, Computer Assisted Web Interview). Es handelt es sich um eine Quotenstichprobe, geschichtet nach Geschlecht, Alter und Bildung ($N$ = 741). Grundgesamtheit waren die Teilnehmer eines Online-

Tabelle 1 Charakteristika der drei Stichproben

| | | Stichprobe 1 | Stichprobe 2 | Stichprobe 3 | Stichprobe 4 |
|---|---|---|---|---|---|
| *Stichprobe* | | | | | |
| Umfang [*N*] | | 539 | 741 | 1.134 | 56 |
| Art | | Quote | Quote | Zufall | Gelegenheit |
| Modus | | CAPI, Papier | CAWI | CAPI, CASI | Papier |
| *Zusammensetzung* | | | | | |
| Geschlecht [% Frauen] | | 52,5% | 51,8% | 55,6% | 82% |
| Alter [M(SD)] | | 47.2 (15.2) | 48.3 (13.0) | 53.3 (18.4) | 24.4 (5.9) |
| Bildung | ≤ 9 Jahre | 44,7% | 40,1% | 37,2% | - |
| | 10 Jahre | 30,2% | 29,1% | 37,0% | - |
| | ≥ 11 Jahre | 23,7% | 30,8% | 25,8% | 100% |

*CAPI = Computer Assisted Personal Interview, CAWI = Computer Assisted Web Interview, CASI = Computer Assisted Self Interview, Papier = Papierversion (Selbstausfüller).*

Access-Panels im Alter von 18 Jahren oder älter, die in Deutschland leben. Die Bearbeitungsdauer des Frageborgens betrug im Mittel 23 Minuten (*SD* = 8).

Stichprobe 3 mit *N* = 1.134 Befragungspersonen ist eine Zufallsstichprobe, die repräsentativ für die Wohnbevölkerung in Deutschland über einem Alter von 18 Jahren ist. Sie wurde mithilfe des ADM-Stichprobensystem Face-to-Face (Random Route) der Arbeitsgemeinschaft deutscher Marktforschungsinstitute gezogen. Die Daten wurden im CAPI- und CASI-Modus (Computer Assisted Self Interview) erhoben. Die Erhebung dauerte durchschnittlich 43 Minuten (*SD* = 13).

Stichprobe 4 (*N* = 56) wurde im Rahmen von mehreren Lehrveranstaltungen des psychologischen Instituts der Universität Potsdam erhoben. Die Bearbeitungszeit des Papierfragebogens betrug maximal 2 Minuten.

## 2.2 Material

### 2.2.1 FACES-Datenbank

Als Ausgangspunkt für die Skalenkonstruktion wurde die FACES-Datenbank des Max-Planck-Instituts für Bildungsforschung verwendet (Ebner et al. 2010). FACES ist eine Sammlung von Portraitfotos von 171 Männern und Frauen dreier Altersklassen (junges, mittleres und hohes Alter). Von jeder der 171 Personen gibt es je zwei Fotos mit neutralem, traurigem, angewidertem, ängstlichem, wütendem und glücklichem Gesichtsausdruck. Durch die Vorselektion einer Modellagentur wur-

de sichergestellt, dass die in der Datenbank abgebildeten Personen (fortan auch: Modelle) möglichst durchschnittlich aussehen und keine auffälligen Merkmale wie Tätowierungen oder Piercings aufweisen.

### 2.2.2 AR1

Das AR1 zur Einschätzung der Attraktivität besteht aus einem Item. Zusammen mit einer elfstufigen Antwortskala wird das Bild einer durchschnittlich attraktiven Person des gleichen Geschlechts und der gleichen Altersgruppe wie die zu beurteilende Person dargeboten (Ankerbild). Die Einschätzung der Attraktivität der Zielperson erfolgt relativ zu diesem Vergleichsmaßstab. Die elf Antwortkategorien des AR1 reichen von „viel unattraktiver" (-5) bis „viel attraktiver" (5) mit einer neutralen Mittelkategorie „gleich attraktiv" (0), über der das Ankerbild dargeboten wird (siehe Abbildung 1). Die Instruktion an den Beurteiler lautet: „Bitte schätzen Sie die Attraktivität der Befragungsperson ein. Geben Sie an, wie attraktiv oder unattraktiv der Befragte im Vergleich zu der unten abgebildeten Person ist." (Instruktion für die Fremdeinschätzung; bei einer Selbsteinschätzung muss die Instruktion entsprechend angepasst werden).

Insgesamt gibt es sechs Skalenvarianten, die sich nur im Ankerbild unterscheiden. Jede enthält ein Ankerbild aus der FACES-Datenbank, das dem Geschlecht (männlich oder weiblich) und dem Alter (junges, mittleres oder hohes Alter) der zu beurteilenden Zielperson bestmöglich entspricht. Die Bilder mit den folgenden IDs wurden aus Set A der Datenbank ausgewählt: Nr. 167 als Anker für die Beurteilung von jungen Männern, Nr. 054 für die Beurteilung von jungen Frauen, Nr. 178 für die Beurteilung mittelalter Männer, Nr. 080 für die Beurteilung mittelalter Frauen, Nr. 027 für die Beurteilung älterer Männer und Nr. 133 für die Beurteilung älterer Frauen. Aufgrund des Copyrights können die sechs Ankerbilder, die als Grundlage der Skalenvarianten dienen, in dieser Publikation nicht abgedruckt werden. Sie können über die Onlinepräsenz des Max-Planck-Instituts einfach und kostenlos angefordert werden (http://faces.mpdl.mpg.de). Eine Anleitung dazu findet sich bei Kemper et al. (2012).

### 2.2.3 Fragebögen

Der Fragebogen von Stichprobe 1 enthielt soziodemographische Angaben, verschiedene psychologische und sozialwissenschaftliche Skalen und 36 Bilder von Modellen aus der FACES-Datenbank. Die Attraktivität der Modelle wurde auf einer Skala von „gar nicht attraktiv" (1) bis „sehr attraktiv" (7) beurteilt. Der Fragebogen

Abbildung 1 Attraktivitätsrating 1 mit Beispiel-Ankerbild

| -5 | -4 | -3 | -2 | -1 | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|---|---|
| viel unattraktiver | | | | | gleich attraktiv | | | | | viel attraktiver |

*Anmerkung: Das Ankerbild in Abbildung 1 ist lediglich ein Beispiel. Die Skala darf in dieser Form nicht eingesetzt werden. Alle notwendigen Informationen zur Erstellung der sechs Skalenvarianten des AR1 sind in einem Arbeitsbericht von Kemper et al. (2012) zu finden.*

in Stichprobe 2 enthielt soziodemographische Angaben, verschiedene Skalen zur Erfassung psychologischer Merkmale und 18 Bilder von Modellen, deren Attraktivität mit dem AR1 eingeschätzt werden sollte. Die Fragebogenbatterie in Stichprobe 3 beinhaltete neben dem AR1 umfangreiche soziodemographische Angaben und verschiedene psychologische und sozialwissenschaftliche Maße, z.B. zur Erfassung von Selbstwirksamkeit (AKSU; Beierlein et al. - dieses Heft), kristalliner Intelligenz (BEFKI GC-K; Schipolowski et al. – in diesem Heft), positiver und negativer Affektivität (Menold/Kemper 2013), physischer und psychischer Beeinträchtigung der Gesundheit (Andersen/Mühlbacher/Nübling 2007), Einkommen und Sozialstatus. Das AR1 war zweimal im Fragebogen enthalten. Am Anfang des Interviews (CAPI) nahm der Interviewer eine Einschätzung der Attraktivität der Befragungsperson vor. Am Ende des Interviews folgte ein kurzer Teil des Fragebogens im CASI-Modus. In diesem sollte die Befragungsperson selbst ihre Attraktivität anhand des AR1 einschätzen. Der Fragebogen von Stichprobe 4 enthielt neben dem AR1 zur Bewertung von sechs Modellen nur Fragen nach Alter und Geschlecht der Beurteiler.

## 2.3 Vorgehen und statistische Analysen

Die Entwicklung und Validierung des AR1 erfolgte in drei Schritten: (1) Vorauswahl durchschnittlich attraktiver Modelle aus der FACES-Datenbank, (2) Selektion von Ankerbildern aufgrund quantitativer Daten aus Stichprobe 1 und (3) empirische

Überprüfung der psychometrischen Güte des AR1 anhand der Stichproben 2, 3 und 4.

Für die Konstruktion des AR1 wurden 171 Bilder von Personen mit neutralem Gesichtsausdruck aus der FACES-Datenbank entnommen und deren Attraktivität ermittelt. Ziel war es, für jede mögliche Zielperson einen im Hinblick auf Alter und Geschlecht möglichst repräsentativen Vertreter mittlerer Attraktivität als Ankerbild auszuwählen. Die 171 Bilder der Modelle wurden daher in sechs Gruppen eingeteilt, in junge, mittelalte und ältere Männer und Frauen (jeweils 27-29 Bilder) und vier Beurteilern aus dem GESIS-Projekt Standardskalen vorgelegt. Diese schätzten die Attraktivität der Modelle anhand einer siebenstufigen Skala von „gar nicht attraktiv" bis „sehr attraktiv" ein. Ziel hierbei war eine erste grobe Einschätzung, um Modelle aus dem Bilderpool zu entfernen, die von einer durchschnittlichen Attraktivität abweichen. Die Einschätzungen der Beurteiler wurden aggregiert, um stabilere Schätzungen zu erhalten. Anschließend wurden aus jeder Alter×Geschlecht-Gruppe 10 Bilder selegiert mit ungefähr mittleren Attraktivitätseinschätzungen (nach Standardisierung wurden pro Gruppe diejenigen Modelle mit z-Werten um 0 selegiert) und einem wahrgenommenen Alter im Bereich von 25-30, 45-50, 65-70 Jahren (Alterseinschätzungen berichtet bei Ebner et al. 2010).

Im zweiten Schritt wurden aus den 10 Bildern pro Gruppe weitere vier aufgrund markanter Merkmale nach Expertendiskussion (Projekt Standardskalen) entfernt. Pro Gruppe verblieben sechs Bilder. Die resultierenden 36 Bilder wurden mit Instruktionen und Antwortskala in den Fragebogen von Stichprobe 1 integriert. Für die Auswahl von Ankerbildern durchschnittlich attraktiver Modelle wurde die Stichprobe der Befragungspersonen beziehungsweise Beurteiler nach Alter und Geschlecht in sechs Gruppen unterteilt. Die Attraktivitätseinschätzungen wurden anschließend innerhalb jeder Gruppe gemittelt und aus jeder Gruppe wurde ein Bild mit mittlerer Attraktivitätseinschätzung gewählt. Das heißt, dass zum Beispiel das Ankerbild für die Einschätzung der Attraktivität von älteren Männern anhand der gemittelten Einschätzungen der älteren männlichen Beurteiler aus Stichprobe 1 ausgesucht wurde. Auf diese Weise wurden für alle sechs Alter×Geschlecht-Gruppen Ankerbilder ausgewählt.

Im dritten Schritt wurden zwei unterschiedliche Aspekte der Konstruktvalidität überprüft und die Reliabilität des AR1 geschätzt. Die Konstruktvalidität wurde über Unterschiede der AR1-Mittelwerte zwischen Modellen mit unterschiedlichen Attraktivitätsniveaus und Zusammenhänge des AR1 mit aus der Literatur bekannten Korrelaten von physischer Attraktivität ermittelt. Für jede der sechs Skalenvarianten wurden drei Bilder von Modellen ausgewählt, die von den Beurteilern in Stichprobe 1 als gering (z-Wert ca. -3), mittel (z-Wert ca. 0)

oder hoch attraktiv (z-Wert ca. 3) eingeschätzt wurden. Wenn das AR1 eine valide Einschätzung der Attraktivität erlaubt, dann sollten sich die AR1-Mittelwerte der drei unterschiedlich attraktiven Modelle in einer separaten Stichprobe deutlich voneinander unterscheiden. Die Einschätzung der Modelle aus Stichprobe 1 in „gering", „mittel" oder „hoch attraktiv" sollte sich in den Mittelwerten des AR1 aus Stichprobe 2 widerspiegeln. Wenngleich die Attraktivität in beiden Stichproben mit unterschiedlichen Messkonzepten (absolute vs. relative Einschätzung) erfasst wurde, sollten die Rangreihen der Messwerte dennoch miteinander korrespondieren. Um dies zu prüfen, wurden sechs Varianzanalysen mit Messwiederholung durchgeführt. Messwiederholungsfaktor war die Attraktivität der Modelle (gering, mittel, hoch, klassifiziert nach Stichprobe 1). Ein linearer Trendtest wurde verwendet, um die A-priori-Hypothese steigender AR1-Werte über die drei Attraktivitätsstufen hinweg zu prüfen. Eine weitere Überprüfung der Konstruktvalidität des AR1 fand in Stichprobe 3 statt, indem die Attraktivitätseinschätzungen mit verschiedenen Variablen korreliert wurden, die mit physischer Attraktivität in Zusammenhang stehen. Ziel war hierbei prinzipiell eine Replikation von in der Literatur berichteten Korrelationen. Gleichzeitig sollten aber auch mögliche Unterschiede zwischen Kriteriumskorrelationen von absoluten und relativen Attraktivitätseinschätzungen identifiziert werden. Weiterhin wurde das AR1 in Stichprobe 3 sowohl dem Interviewer als auch der Befragungsperson vorgelegt, um, als weiteren Validitätsbeleg, die Übereinstimmung zwischen Selbst- und Fremdeinschätzung physischer Attraktivität zu ermitteln.

Ein Schätzwert für die Reliabilität der Attraktivitätseinschätzungen mit dem AR1 wurde in Form der Beurteilerübereinstimmung (Interraterreliabilität) für jede der sechs Skalenvarianten in Stichprobe 4 bestimmt. Bei der Bewertung einer Zielperson durch mehrere Beurteiler wird üblicherweise die Konsistenz und/oder die Übereinstimmung der Urteile über die Berechnung statistischer Kennwerte wie z.B. Cronbach $\alpha$ (Cronbach 1951) oder dem Intraklassenkorrelationskoeffizienten (ICC; McGraw/Wong 1996; Shrout/Fleiss 1979) ermittelt, um sicher zu stellen, dass den Einschätzungen ein gemeinsamer Attraktivitätsstandard zugrunde liegt. Je stärker dies der Fall ist, desto ähnlicher fallen die Attraktivitätsurteile auch aus. Im Gegensatz zu Konsistenzmaßen wie Cronbach $\alpha$ berücksichtigen Maße der Übereinstimmung auch systematische Niveauunterschiede zwischen den Beurteilern. Im Vergleich zur hohen Konsistenz von Attraktivitätsurteilen ($r$ = .85 – .94, Rubenstein et al. 2002), liegen Kennwerte der Übereinstimmung häufig auf einem niedrigen bis mittleren Niveau (ICCs = .26 – 34, Hassebrauck 1993; .30 – .50, Thornhill/Gangestad 1999). Die offensichtliche Diskrepanz dieser beiden Ansätze wird bei Hönekopp (2006) und Hassebrauck (1993) kritisch diskutiert. Um die absolute

Übereinstimmung der Beurteiler zu bestimmen, wurde nach den Empfehlungen von McGraw und Wong (1996) der Intraklassenkorrelationskoeffizient (ICC[A, 1]) berechnet. Dazu beurteilten alle Probanden in Stichprobe 4 die Attraktivität von sechs Modellen anhand des AR1. Die Modelle wurden als Portraitfotos oberhalb des AR1 mit der Instruktion präsentiert, die Attraktivität der Modelle anhand des AR1 einzuschätzen.

# 3 Ergebnisse

## 3.1 Deskriptive Statistiken

Die mittlere Attraktivitätseinschätzung der Zielpersonen in der bevölkerungsrepräsentativen Stichprobe 3 liegt bei 1.66 (SD = 2.14). Die Interviewer schätzten die Zielpersonen somit im Mittel als mehr als eine Skalenstufe attraktiver ein als die AR1-Ankerfotos. In Tabelle 2 sind deskriptive Statistiken für das AR1 nach Alter, Geschlecht und Bildung aufgeführt, um den Anwendern einen Vergleich der AR1-Werte aus ihrer Untersuchung mit denen relevanter Subgruppen aus einer bevölkerungsrepräsentativen Zufallsstichprobe, zum Beispiel von Männern oder Frauen, von Personen mit unterschiedlicher Schulbildung oder unterschiedlichen Alters, zu ermöglichen.

## 3.2 Reliabilität

Die Beurteilungen der sechs Modelle in Stichprobe 4 fielen sehr konsistent aus (Cronbach $\alpha$ = .99). Die Reliabilität des AR1 wurde anhand der absoluten Beurteilerübereinstimmung geschätzt. Diese wurde über die Berechnung des Intraklassen-Korrelationskoeffizienten (ICC[A, 1], McGraw/Wong 1996) bestimmt. Sie fällt mit ICC = .65 in den Bereich angemessener Reliabilität (vgl. Shrout 1998) und liegt somit über den üblicherweise in der Literatur berichteten Kennwerten (siehe Diskussion).

## 3.3 Validität

Die Ergebnisse der AR1-Einschätzungen in Stichprobe 2 von Modellen mit geringer, mittlerer und hoher Attraktivität (eingeschätzt in Stichprobe 1) und Signifikanztests der Varianzanalysen sind in Tabelle 3 zu finden. Bei allen sechs Skalenvarianten unterschieden sich die Attraktivitätseinschätzungen der drei Modelle signifikant voneinander (alle $Fs_{[2,1480]} \geq 205.99$, $ps < .001$, $\eta_p^2 \geq .22$). Modelle, die

Tabelle 2 Deskriptive Statistiken (Stichprobe 3)

| | | AR1-Werte (Fremdeinschätzung) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 18-35 Jahre | | 36-65 Jahre | | >65 Jahre | | Gesamt | |
| Geschlecht | Bildung | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Männlich | gering | 1.23 | 1.80 | 1.05 | 2.15 | 0.83 | 1.95 | 0.97 | 2.01 |
| | mittel | 0.75 | 2.32 | 1.57 | 1.99 | 0.80 | 1.53 | 1.21 | 1.96 |
| | hoch | 1.92 | 1.95 | 2.51 | 1.71 | 1.20 | 1.79 | 2.05 | 1.86 |
| | Gesamt | 1.38 | 2.07 | 1.67 | 2.05 | 0.89 | 1.82 | 1.34 | 2.00 |
| Weiblich | gering | 2.11 | 2.49 | 1.59 | 2.47 | 1.31 | 1.98 | 1.49 | 2.23 |
| | mittel | 2.16 | 2.12 | 2.14 | 2.30 | 1.38 | 1.89 | 2.02 | 2.20 |
| | hoch | 2.44 | 2.07 | 2.43 | 2.26 | 1.77 | 1.82 | 2.34 | 2.13 |
| | Gesamt | 2.26 | 2.14 | 2.06 | 2.35 | 1.39 | 1.94 | 1.92 | 2.21 |
| Gesamt | gering | 1.60 | 2.14 | 1.33 | 2.33 | 1.09 | 1.98 | 1.23 | 2.14 |
| | mittel | 1.76 | 2.26 | 1.92 | 2.20 | 1.08 | 1.72 | 1.71 | 2.15 |
| | hoch | 2.23 | 2.03 | 2.46 | 2.01 | 1.44 | 1.81 | 2.21 | 2.01 |
| | Gesamt | 1.92 | 2.15 | 1.89 | 2.23 | 1.14 | 1.89 | 1.66 | 2.14 |

*N = 1.134*

in Stichprobe 1 als „gering attraktiv" eingeschätzt wurden, zeigten in Stichprobe 2 geringere Mittelwerte im Vergleich zu Modellen, die als „mittel" oder „hoch attraktiv" eingeschätzt wurden. Signifikante lineare Trends der Mittelwerte über die drei Attraktivitätsstufen konnten für alle Skalenvarianten beobachtet werden (alle $F$s$_{[1,740]} \geq 260.76$, $p$s $< .001$, $\eta_p^2 \geq .31$). Ein quadratischer Trend konnte nur für die Skala „weiblich, alt" mehr Varianz aufklären als der lineare Trend ($F_{[1,740]} = 360.83$, $p < .001$, $\eta_p^2 = .33$). Für fünf der sechs Skalenvarianten steigen die Mittelwerte des AR1 also mit den Attraktivitätsstufen von „gering" über „mittel" bis „hoch attraktiv" an. Einzige Ausnahme war die Skalenvariante zur Einschätzung von älteren Frauen. Hier korrespondierte die Höhe der Mittelwerte in der AR1 nicht mit den Stufen „mittel" und „hoch attraktiv". Die Stufe „hoch attraktiv" wies einen niedrigeren Mittelwert als die Stufe „mittel" auf. Da die übrigen Befunde klare Übereinstimmungen der Messwertreihen zeigen, könnte dieser Effekt durch die spezifische Auswahl der Modelle für die Gruppe „weiblich, alt" erklärt werden. Die praktische Bedeutsamkeit der Effekte der beobachteten Mittelwertsunterschiede und des linearen Anstiegs der Mittelwerte ist als groß einzustufen ($\eta_p^2 > .14$, nach Cohen, 1988). Demnach erlaubt das AR1 eine valide Erfassung unterschiedlicher Grade der Attraktivität von Befragungspersonen.

Tabelle 3 Mittelwerte und Standardabweichungen der AR1-Einschätzungen in Stichprobe 2

| | Attraktivität | | | Trend (linear) | | Trend (quadratisch) | |
|---|---|---|---|---|---|---|---|
| Gruppe | gering | mittel | hoch | $F_{[1,740]}$ | $\eta_p^2$ | $F_{[1,740]}$ | $\eta_p^2$ |
| Männlich jung | -0.91 (2.00) | 0.11 (1.94) | 1.07 (2.11) | 349.96*** | .32 | 0.24 | .00 |
| Weiblich, jung | -0.50 (2.07) | 1.03 (2.00) | 2.20 (2.21) | 590.30*** | .44 | 15.79* | .01 |
| Männlich, mittelalt | -1.15 (1.93) | -0.69 (1.89) | 2.01 (2.17) | 798.92*** | .52 | 214.79*** | .23 |
| Weiblich, mittelalt | -0.96 (1.98) | -0.43 (1.73) | 1.12 (2.19) | 385.16*** | .34 | 56.27*** | .07 |
| Männlich, alt | -1.63 (2.22) | -0.16 (1.96) | 0.33 (2.13) | 337.18*** | .31 | 45.83*** | .06 |
| Weiblich, alt | -0.81 (1.68) | 1.46 (1.95) | 0.64 (1.81) | 260.76*** | .26 | 360.83*** | .33 |

*N= 741. F = Prüfstatistik des F-Tests. $\eta_p^2$ = Effektstärke (partielles Eta-Quadrat). * = p < .05, *** = p < .001.*

Um weitere Belege für die Konstruktvalidität des AR1 zu finden, wurde anhand von Stichprobe 3 versucht, aus der Fachliteratur bekannte typische Korrelationen von Attraktivität zu replizieren und einen Vergleich zwischen den anhand des AR1 berechneten Validitätskoeffizienten mit auf relativen Attraktivitätsmaßen beruhenden Korrelationen vorzunehmen. Die Validitätskoeffizienten sind in Tabelle 4 zu finden. Die praktische Bedeutsamkeit der im Folgenden berichteten Validitätskoeffizienten wird nach den Richtlinien von Cohen (1992) vorgenommen: kleiner Effekt ($r$ = .10), mittlerer Effekt ($r$ = .30), starker Effekt ($r$ = .50). Die höchste Korrelation findet sich erwartungsgemäß zwischen der Selbst- und Fremdeinschätzung der Befragungspersonen anhand des AR1 ($r$ = .46, $p$ < .01). Diese Korrelation liegt im oberen Bereich der üblicherweise in der Literatur berichteten Übereinstimmungen zwischen Fremd- und Selbsteinschätzungen physischer Attraktivität. Feingold (1992) berichtet z.B. in einer metaanalytischen Untersuchung eine mittlere Übereinstimmungen von $r$ = .24. Des Weiteren finden sich signifikante, wenn auch nominell kleine Zusammenhänge der AR1-Werte mit soziodemographischen Variablen. So sind als attraktiver bewertete Befragungspersonen jünger, verfügen über ein höheres Einkommen, eine bessere Bildung, ein größeres soziales Netzwerk und sind häufiger in einer festen Beziehung. Eine negative Korrelation physischer Attraktivität mit dem Alter wird häufig berichtet (z.B. Henss 1991; Jones/Hill, 1993; Teuscher/Teuscher 2007). Die mit dem AR1 erhobenen Urteile zeigen dabei einen schwächeren Zusammenhang mit dem Alter als häufig in der Literatur angegeben ($r$ = -.18, $p$ < .01). Henss (1991) fand z.B. in zwei Experimenten Korrelationen zwischen physischer Attraktivität und Alter im Bereich von $r$ = -.33 – -.83 bei Frauen

und $r = .00 - -.69$ bei Männern. Der hier gefundene Zusammenhang mit dem Einkommen ($r = .11$, $p < .05$) deckt sich mit einer Reihe in der Literatur berichteter Korrelationen physischer Attraktivität mit diversen Kriterien des beruflichen und akademischen Erfolgs. Pfeifer (2011) sowie Roszell et al. (2001) fanden dabei ähnliche Effektstärken für den Zusammenhang von Attraktivität und Einkommen ($r = .03 - .08$). Die Korrelation mit dem Bildungsniveau ($r = .18$, $p < .01$) bestätigt die Ergebnisse von Umberson und Hughes (1987), die ebenfalls einen linearen Zusammenhang dieses Merkmals mit physischer Attraktivität berichten. Vor dem Hintergrund, dass als attraktiver wahrgenommene Personen stärker ausgeprägte soziale Fertigkeiten berichten und in sozialen Situationen weniger befangen sind (Feingold 1992), erscheint es auch plausibel, dass sie über ein größeres soziales Netzwerk verfügen ($r = .11$, $p < .01$). Weiterhin sind attraktive Menschen eher in einer Partnerschaft als weniger attraktive ($r = .17$, $p < .01$). Die Höhe dieser Korrelation ist vergleichbar mit den Ergebnissen, die bei Townsend und Levy (1990) für verschiedene Beziehungstypen angegeben werden ($r = .15 - .27$). Frauen wurden generell als attraktiver beurteilt ($r = .13$, $p < .01$). Übereinstimmend mit Ergebnissen der Attraktivitätsforschung korrelieren die AR1-Werte mit kristalliner Intelligenz ($r = .20$, $p < .01$). Zebrowitz et al. (2002) berichten analog, abhängig vom Alter der beurteilten Personen, Korrelationen von $r = .11 - .26$ zwischen Attraktivität und Intelligenz. Die Interviewer schrieben attraktiveren Personen außerdem einen höheren sozialen Status zu ($r = .22$, $p < .01$). Tatsächliche Unterschiede bezüglich sozialer Kompetenz und Dominanz zwischen unterschiedlich attraktiven Menschen werden in diversen Studien berichtet (z.B. Eagly et al. 1991; Feingold 1992; Langlois et al. 2000). Erwartungsgemäße Zusammenhänge finden sich auch bezüglich der Variablen im Bereich Gesundheit und Lebensqualität. Physische Attraktivität korreliert positiv mit der allgemeinen Lebenszufriedenheit ($r = .19$, $p < .01$) und positiver Affektivität ($r = .29$, $p < .01$). Vergleichbare Zusammenhänge werden z.B. bei Umberson und Hughes (1987) für positiven Affekt ($\beta = .10$) und Lebenszufriedenheit ($\beta = .06$) berichtet. Negative Zusammenhänge finden sich entsprechend mit selbsteingeschätzten physischen und psychischen Beeinträchtigungen der Gesundheit ($r = -.18$ bzw. $-.12$, $p$s $< .01$) sowie negativer Befinden ($r = -.11$, $p < .01$). Dies deckt sich mit den Ergebnissen von Shackelford und Larsen (1999), die u.a. Zusammenhänge von physischer Attraktivität mit besserer kardiovaskulärer Gesundheit ($r = .15$) und seltenerem Auftreten von Kopfschmerzen ($r = -.17$) berichten. Weiterhin besteht ein positiver Zusammenhang mit der allgemeinen Selbstwirksamkeit ($r = .21$, $p < .01$), ähnlich den in der Literatur berichteten stärker ausgeprägten internalen Kontrollüberzeugungen attraktiver Menschen (Feingold 1992).

Tabelle 4 Validitätskoeffizienten des AR1 in Stichprobe 3

| | AR1 (Fremdeinschätzung) |
|---|---|
| *Attraktivität* | |
| Selbsteinschätzung | .46** |
| *Soziodemographische Variablen* | |
| Alter | -.18** |
| Geschlecht | .13** |
| Einkommen [1] | .11* |
| Bildung | .18** |
| Größe des sozialen Netzwerks (Anzahl Personen) | .11** |
| Partnerschaft ja/nein [2] | .17** |
| Länge der Beziehung in Jahren[2] | -.14** |
| *Status und Erfolg* | |
| Anerkennung durch Vorgesetzten [3] | .08 |
| Zugeschriebener sozialer Status | .22** |
| *Kognitive Fähigkeiten* | |
| Kristalline Intelligenz | .20** |
| *Gesundheit und Lebensqualität* | |
| Positive Affektivität | .29** |
| Negative Affektivität | -.11** |
| Beeinträchtigung physisch | -.18** |
| Beeinträchtigung psychisch | -.12** |
| Lebenszufriedenheit | .19** |
| *Selbstwirksamkeit* | |
| Allgemeine Selbstwirksamkeit | .21** |

*N = 1097-1134. Geschlecht = männlich (1), weiblich (2). Bildung = gering (1), mittel (2), hoch (3). Partnerschaft = (1) nein, (2) ja. [1]N = 541, [2]N = 534, [3]N = 462. * = p < .05, ** = p < .01.*

# 4 Diskussion

Durch ihre besondere Rolle in Prozessen der sozialen Kognition, z.B. der Eindrucksbildung, und ihrer Implikationen für menschliches Erleben und Verhalten ist die physische Attraktivität für eine Vielzahl sozialer Phänomene relevant. Vermehrt wird diese daher auch in Umfragen erfasst. Die weit verbreitete Messung über absolute Globalurteile auf Ratingskalen weist allerdings methodische Schwächen auf. In der hier beschriebenen Skalenentwicklung wird erstmals ein relatives Messkon-

zept umgesetzt, um diesen methodischen Schwächen Rechnung zu tragen und die Standardisierung der Attraktivitätseinschätzungen zu verbessern.

Das AR1 wurde in mehreren Schritten anhand von vier Stichproben entwickelt und validiert. In einem mehrstufigen Verfahren wurden zunächst Skalenvarianten für die Attraktivitätsbeurteilung von jungen, mittelalten und älteren Männern und Frauen erstellt und anschließend auf ihre psychometrische Güte hin überprüft. Dafür wurden Attraktivitätseinschätzungen mit dem AR1 mit Attraktivitätseinschätzungen, die mit einem üblicherweise eingesetzten absoluten Messverfahren erhoben wurden, verglichen. Weiterhin wurden anhand einer bevölkerungsrepräsentativen Zufallsstichprobe Validitätskoeffizienten ermittelt und aus der Fachliteratur bekannte Korrelate der physischen Attraktivität anhand des AR1 repliziert. In einer weiteren Stichprobe wurde die Reliabilität des AR1 anhand der Beurteilerübereinstimmung bestimmt.

Im Folgenden wird die psychometrische Güte des AR1 diskutiert: Objektivität, Reliabiliät und Konstruktvalidität. Die Durchführungsobjektivität hängt bei einem Face-to-face-Interview von dem Interviewer ab, der die Daten erhebt. Sie ist gegeben, wenn dieser sich bei der Messung an die Instruktionen hält. Bei entsprechend geschulten Interviewern ist die Durchführungsobjektivität also üblicherweise gewährleistet (Rammstedt 2010). Auswertungs- und Interpretationsobjektivität sind durch das Messkonzept des AR1 ebenfalls gewährleistet.

Die von verschiedenen Beurteilern mit dem AR1 erhobenen Attraktivitätsurteile der gleichen Personen zeigten ein hohes Maß an Übereinstimmung. Die anhand von Stichprobe 4 berechnete Interraterreliabilität fällt dabei höher aus, als dies in der Literatur für Attraktivitätsbeurteilungen üblicherweise berichtet wird (vgl. Thornhill/Gangestad 1999). Nach den Kriterien von Shrout (1998) zu Anforderungen an Instrumente der klinischen Diagnostik liegen die in dieser Studie ermittelten Werte im Bereich angemessener Interraterreliabilität. Dieses Ergebnis könnte einen ersten Hinweis darauf darstellen, dass den mit dem AR1 erhobenen Attraktivitätsurteilen in stärkerem Maß ein geteilter Bewertungsmaßstab zugrunde liegt, als dies bei herkömmlich verwendeten, absoluten Messinstrumenten der Fall ist.

Im Rahmen der Validitätsanalysen zeigte sich, dass die AR1-Werte von Personen, die in Stichprobe 1 als „gering", „mittel" oder „hoch attraktiv" eingeschätzt wurden, linear über die „Attraktivitätsstufen" ansteigen, d.h. Personen, die in Stichprobe 1 als „gering attraktiv" eingeschätzt wurden, zeigten in Stichprobe 2 signifikant geringere Mittelwerte im Vergleich zu Personen, die als „mittel" oder „hoch attraktiv" eingeschätzt wurden, und als „hoch attraktiv" eingeschätzte Personen zeigten im Mittel die höchsten AR1-Werte im Vergleich zu Personen, die als „gering" oder „mittel" attraktiv eingeschätzt wurden. Das bedeutet, dass die mit

dem AR1 erhobenen, relativen Attraktivitätswerte grundsätzlich mit standardmäßig eingesetzten absoluten Ratingverfahren vergleichbar sind.

Die empirischen Validitätskoeffizienten des AR1 spiegeln die erwarteten und aus der Literatur bekannten Beziehungen des Konstrukts physische Attraktivität angemessen wider. Die Korrelation zwischen der Fremdeinschätzung der Interviewer und der Selbsteinschätzung der Befragungsperson fiel dabei höher aus, als dies in der Attraktivitätsforschung üblicherweise berichtet wird. Weitere mit der Fachliteratur konsistente Korrelationen von geringer bis mittlerer Stärke wurden mit dem Alter, mit Indikatoren der Gesundheit und der Lebenszufriedenheit, sozioökonomischen Erfolgsvariablen, sozialen Ressourcen und dem zugeschriebenen sozialen Status der Befragungsperson gefunden. Die Ergebnisse der Korrelationsanalysen lassen sich, wie im Folgenden dargestellt, gut in die bestehende Attraktivitätsforschung integrieren.

Wie bei Attraktivitätsbeurteilungen häufig berichtet (z.B. Henss 1991), zeigen die mit dem AR1 erhobenen Urteile einen negativen Zusammenhang mit dem Alter der zu bewertenden Person. Ältere Menschen werden als weniger attraktiv eingeschätzt. Im Vergleich zu in der Literatur berichteten Korrelationen mit absoluten Messverfahren fällt der Einfluss des Alters der Zielperson auf die AR1-Werte allerdings geringer aus. Dieser vergleichsweise geringe Zusammenhang der AR1-Werte mit dem Alter der Zielpersonen lässt den vorläufigen Schluss zu, dass das hier verwendete, relative Messkonzept tatsächlich eine weniger durch Beurteilermerkmale verzerrte Messung physischer Attraktivität erlaubt als die üblicherweise verwendeten absoluten Urteile.

Die Korrelationen von physischer Attraktivität mit dem Einkommen, dem Bildungsniveau und der Zuschreibung von sozialem Status fielen in Richtung und Stärke erwartungsgemäß aus und zeichnen ein Bild von Attraktivität als einem generellen Prädiktor sozioökonomischen Erfolgs. Feingold (1992; siehe auch Kalick 1988) bezeichnet physische Attraktivität als Statuscharakteristikum. Ein entsprechender Effekt auf die Eindrucksbildung findet sich auch im Rahmen der vorliegenden Untersuchung. Die Interviewer ordneten attraktivere Personen signifikant häufiger in eine statushöhere Gesellschaftsschicht ein. Entgegen den Erwartungen fand sich in den hier präsentierten Daten allerdings kein Zusammenhang mit der empfundenen Anerkennung durch Vorgesetze. Eine positive Diskriminierung aufgrund der Attraktivität (vgl. Morrow et al. 1990) kann dadurch allerdings nicht ausgeschlossen werden, da sie sich evtl. primär in Einstellungsentscheidungen ausdrückt.

Die gefundene Korrelation mit Intelligenz lässt sich gut in die Befunde von Zebrowitz et al. (2002) einordnen, die ähnlich starke Zusammenhänge fanden.

Anzumerken ist hier, dass Zebrowitz et al. auf vollständige IQ-Werte (Stanford-Binet Test und Wechsler Adult Intelligence Scales – Revised) zurückgreifen konnten, während in der vorliegenden Studie aus ökonomischen Gründen nur eine Intelligenzkomponente, die kristalline Intelligenz, betrachtet wurde. Dadurch ist nicht auszuschließen, dass die Varianz der Kriteriumsvariable Intelligenz in dieser Studie eingeschränkt war. Untersuchungen mit einer elaborierteren Intelligenzdiagnostik könnten möglicherweise noch stärkere Korrelationen mit dem AR1 finden. Dass Attraktivität offenbar ein generelles Kriterium der Partnerwahl ist, zeigt sich auch in der hier betrachteten Stichprobe. Attraktive Menschen sind häufiger in einer Partnerschaft als weniger Attraktive. Während attraktivere Menschen häufiger in Beziehungen sind als weniger attraktive, gaben sie allerdings eine geringere Beziehungsdauer an.

Die mit dem AR1 gemessene physische Attraktivität steht weiterhin erwartungsgemäß in Zusammenhang mit mehreren Affekt- und Gesundheitsvariablen. Dabei berichten als attraktiver eingeschätzte Menschen positivere Affektivität, eine höhere Lebenszufriedenheit und weniger psychische und physische Beeinträchtigungen. Diese Befunde decken sich mit Ergebnissen anderer Studien, die nahelegen, dass Attraktivität mit positiver Affektivität und subjektiv wahrgenommener Gesundheit korreliert, aber auch mit physischer Gesundheit (Umberson/Hughes 1987; Feingold 1992; Shackelford/Larsen 1999).

In der metaanalytischen Studie von Feingold (1992) zeigt sich auch, dass attraktivere Menschen stärkere Kontrollüberzeugungen, also den Glauben, dass sie ihr Leben selbst kontrollieren, sowie stärker ausgeprägtes Selbstwertgefühl berichten. In der vorliegenden Studie konnte eine positive Korrelation mit dem Konstrukt allgemeine Selbstwirksamkeit festgestellt werden. Das Konzept Selbstwirksamkeit ist in seiner allgemeinen, situationsunspezifischen Form vergleichbar mit dem Konstrukt der internalen Kontrollüberzeugungen. Attraktive Menschen sind im Vergleich zu weniger attraktiven eher der Überzeugung, dass ihre eigenen Handlungen einen starken Einfluss auf wichtige Aspekte ihres Lebens haben.

Zusammenfassend lässt sich feststellen, dass die AR1-Werte in erwarteter Weise mit verschiedenen aus der Attraktivitätsforschung bekannten Variablen korrelieren. Die Höhe der Validitätskoeffizienten lag in den meisten hier betrachteten Fällen in den in der Literatur berichteten Bereichen. Eine Ausnahme stellt die hohe Korrelation von Fremd- und Selbsturteilen dar. Des Weiteren scheinen die AR1-Werte weniger abhängig vom Geschlecht der Zielpersonen zu sein, als dies an anderer Stelle für Attraktivitätsurteile berichtet wird. Wir werten diese Befunde als erste Hinweise auf die Nützlichkeit des im AR1 verwendeten relativen Messkonzepts im Vergleich zu absoluten Attraktivitätsmaßen.

Die berichteten Befunde dieser Studie legen nahe, dass sich anhand des AR1 das Merkmal physische Attraktivität adäquat messen lässt. Allerdings unterliegt die vorliegende Untersuchung und das Messkonzept des AR1 gewissen Einschränkungen. Durch die Verwendung eines Ankerbilds mit dem Gesicht als Grundlage für die relative Einschätzung beziehen wir die physische Attraktivität einer Befragungsperson ausschließlich auf ihr Gesicht. Dies ist in gewisser Weise eine reduktionistische Sichtweise des Konstrukts, deckt sich aber mit Befunden aus der Attraktivitätsforschung, nach denen die Attraktivität des Gesichts der stärkste Prädiktor einer globalen Einschätzung der physischen Attraktivität ist (Mueser et al. 1984; Peters et al. 2007). Außerdem kann aufgrund der aktuellen Daten nicht abschließend beantwortet werden, ob das relative Messkonzept der AR1 tatsächlich klassischen absoluten Maßen überlegen ist. Dazu müssten systematische Vergleiche der beiden Messkonzepte durchgeführt werden. Zukünftige Studien könnten z.B. die beiden Faktoren Messkonzept und Merkmale der Beurteiler (quasi-)experimentell variieren. Da in der vorliegenden Untersuchung eine vergleichsweise niedrige Korrelation von physischer Attraktivität und Alter gefunden wurde, könnte sich das AR1 auch in einer solchen experimentellen Studie als weniger anfällig für Beurteilermerkmale erweisen. Ein möglicher Einwand gegen die grundsätzliche Relevanz des hier betrachteten Merkmals für die sozialwissenschaftliche Umfrageforschung könnte sich auf die Größe der gefundenen Validierungskoeffizienten beziehen. Rein nominell sind diese mit aufgeklärten Varianzanteilen von maximal acht Prozent als klein zu bewerten (Cohen 1988). Auf einer theoretischen Ebene betrachtet, erscheinen die gefundenen Effekte allerdings realistisch und sind vergleichbar mit Ergebnissen aus metaanalytischen Untersuchungen (Eagly et al. 1991; Feingold 1992; Langlois et al. 2000). Kriterien wie Erfolg im Beruf, sozialer Status und die psychische und physische Gesundheit können nur schwerlich vollständig über einige wenige Variablen erklärt werden. Trotzdem bedeutet dies nicht, dass kleine Einflussgrößen ignoriert werden sollten, da auch sie praktisch relevant sein können (vgl. Abelson 1985; Rosenthal 1990). Frieze et al. (1991) berichten z.B. eine Einkommensdifferenz von $2.200 pro Jahr zwischen hoch und niedrig attraktiven Männern unter Kontrolle des Bildungsniveaus.

Die Ergebnisse der Überprüfung der psychometrischen Qualität des AR1 sprechen dafür, dass das AR1 physische Attraktivität reliabel, valide und ökonomisch abbildet. Damit ist das AR1 optimal für den Einsatz in Erhebungen mit beschränkten zeitlichen oder monetären Ressourcen, z.B. in sozialwissenschaftlichen oder gesundheitswissenschaftlichen Umfragen, geeignet.

## Literatur

Abelson, R. P., 1985: A variance explanation paradox: When a little is a lot. Psychological Bulletin 97: 129–133.

Andersen, H. H., A. Mühlbacher u. M. Nübling, 2007: Die SOEP-Version des SF 12 als Instrument gesundheitsökonomischer Analysen. SOEP Papers on Multidisciplinary Panel Data Research, 6. Berlin: DIW.

Asendorpf, J., 2011: Persönlichkeitspsychologie. Heidelberg: Springer.

Bartlett, J. C. und A. Fulton, 1991: Familiarity and recognition of faces in old age. Memory & Cognition 19: 229-238.

Bernstein, N., 1982: Psychological results in burns: The damaged self-esteem. Clinics in Plastic Surgery 9: 337-346.

Bierhoff, H. W., 2000: Sozialpsychologie. Ein Lehrbuch. Stuttgart: Kohlhammer.

Clifford, M. M. und E. Walster, 1973: The effect of physical attractiveness on teacher expectations. Sociology of Education 46: 248-258.

Cohen, J., 1992: A power primer. Psychological Bulletin 112: 155-159.

Cohen, J., 1988: Statistical power analysis for the behavioral sciences. Hillsdale: Erlbaum Associates.

Cronbach, L. J., 1951: Coefficient alpha and the internal structure of tests. Psychometrika 16: 297-334.

Dijkstra, P. und B. P. Buunk, 2001: Sex differences in the jealousy-evoking nature of a rival's body build. Evolution and Human Behavior 22: 335-341.

Eagly, A. H., R. D. Ashmore, M. G. Makhijani und L. C. Longo, 1991: What is beautiful is good, but...: A meta-analytic review of research on the physical attractiveness stereotype. Psychological Bulletin 110: 109-128.

Ebner, N. C., M. Riediger und U. Lindenberger, 2010: FACES: A database of facial expressions in young, middle-aged, and older women and men: Development and validation. Behavior Research Methods 42: 351-362.

Feingold, A., 1992: Good-looking people are not what we think. Psychological Bulletin 111: 304-341.

Fink, B., K. Grammer und R. Thornhill, 2001: Human (Homo sapiens) facial attractiveness in relation to skin texture and color. Journal of Comparative Psychology 115: 92-99.

Fleischer-Peters, A. und J. Margraf-Stiksrud, 1996: Auswirkungen von Normabweichungen auf die Psyche. S. 131-152 in: H.G. Sergl (Hrsg.): Psychologie und Psychosomatik in der Zahnheilkunde. München: Urban & Schwarzenberg.

Frederick, D. A. und M. G. Haselton, 2007: Why is muscularity sexy? Tests of the fitness indicator hypothesis. Personality and Social Psychology Bulletin 33: 1167–1183.

Frieze, I. H., J. E. Olson und J. Russell, 1991: Attractiveness and income for men and women in management. Journal of Applied Social Psychology 21: 1039-1057.

Gallup, G. R. und D. A. Frederick, 2010: The science of sex appeal: An evolutionary perspective. Review of General Psychology 14: 240-250.

Hamermesh, D. S. und J. E. Biddle, 1994: Beauty and the labor market. American Economic Review 84: 1174-1194.

Hassebrauck, M., 1993: Die Beurteilung der physischen Attraktivität. S. 29-59 in: M. Hassebrauck und R. Niketta (Hrsg.): Physische Attraktivität. Göttingen: Hogrefe.

Heimes, H. und C. J. Kemper, 2001: Beeinflusst die Persönlichkeit unsere Reaktion auf physische Attraktivität. Unveröffentlichte Semesterarbeit. Philipps-Universität Marburg.

Henss, R., 1991: Perceiving age and attractiveness in facial photographs. Journal of Applied Social Psychology 21: 933-946.

Hönekopp, J., 2006: Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. Journal of Experimental Psychology – Human Perception and Performance 32: 199-209.

Hosoda, M., E. F. Stone-Romero und G. Coats, 2003: The effects of physical attractiveness on job-related outcomes: A meta-analytical analysis of experimental studies. Personnel Psychology 56: 431-462.

Houston, V. und R. Bull, 1994: Do people avoid sitting next to someone who is facially disfigured? European Journal of Social Psychology 24: 279-284.

Jokela, M., 2009: Physical attractiveness and reproductive success in humans: Evidence from the late 20th century United States. Evolution and Human Behavior 30: 342-350.

Jones, D. und K. Hill, 1993: Criteria of facial attractiveness in five populations. Human Nature 4: 271-296.

Judge, T. A., C. Hurst und L. S. Simon, 2009: Does it pay to be smart, attractive, or confident (or all three)? Relationships among general mental ability, physical attractiveness, core self-evaluations, and income. Journal of Applied Psychology 94: 742-755.

Kalick, S. M., 1988: Physical attractiveness as a status cue. Journal of Experimental Social Psychology 24: 469-489.

Kanazawa, S. und J. L. Kovar, 2004: Why beautiful people are more intelligent. Intelligence 32: 227-243.

Kemper, C. J., J. Lutz, J. Margraf-Stiksrud, C. Beierlein, A. Kovaleva und B. Rammstedt, 2012: Eine Ein-Item-Skala zur Einschätzung von Attraktivität: Das Attraktivitätsrating (AR1). GESIS Working Papers2012/24. GESIS – Leibniz Institut für Sozialwissenschaften. Mannheim. http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_arbeitsberichte/WorkingPapers_2012-24.pdf (27.05.2013).

Kenrick, D. T., und S. E. Gutierrez, 1980: Contrast effects and judgments of physical attractiveness: When beauty becomes a social problem. Journal of Personality and Social Psychology 38: 131-140.

Kenrick, D. T., S. E. Gutierres und L. L. Goldberg, 1989: Influence of popular erotica on judgments of strangers and mates. Journal of Experimental Social Psychology 25: 159-167.

Langlois, J. H., L. Kalakanis, A. J. Rubenstein, A. Larson, M. Hallam und M. Smoot, 2000: Maxims of myths of beauty? A meta-analytic and theoretical review. Psychological Bulletin 126: 390-423.

Lorenz, K., 1943: Die angeborenen Formen möglicher Erfahrung. Zeitschrift für Tierpsychologie 5: 235-409.

Margraf-Stiksrud, J., 1991: Die Funktion des Gesichts für die psychische Entwicklung des Menschen. S. 11-18 in: H.G. Sergl und H. Müller-Fahlbusch (Hrsg.): Jahrbuch der Psychologie und der Psychosomatik in der Zahlheilkunde, Bd. 2. Berlin: Quintessenz.

McGraw, K. O. und S. P. Wong, 1996: Forming inferences about some intraclass correlation coefficients. Psychological Methods 1: 30-46.

Menold, N. und C. Kemper, 2013. Verbal or numerical labels? And how many categories? The impact of rating scale formats on quality metrics of multi-item sets. Manuscript in preparation.

Morrow, P. C., J. C. McElroy, B. G. Stamper und M. A. Wilson, 1990: The effects of physical attractiveness and other demographic characteristics on promotion decisions. Journal of Management 16: 723-736.

Mueser, K. T., B. W. Grau, S. Sussman und A. J. Rosen, 1984: You're only as pretty as you feel: Facial expression as a determinant of physical attractiveness. Journal of Personality and Social Psychology 46: 469-478.

Nedelec, J. L. und K. M. Beaver, 2011: Beauty is in the sex of the beholder: An examination of the effects of interviewer characteristics on assessments of respondent attractiveness. Personality and Individual Differences 51: 930-934.
Peters, M., G. Rhodes und L. W. Simmons, 2007: Contributions of the face and body to overall attractiveness. Animal Behaviour 73: 937-942.
Pfeifer, C., 2011: Physical attractiveness, employment, and earnings. Discussion Paper No. 5664. Leuphana University Lüneburg.
Rammstedt, B., 2010: Reliabilität, Validität, Objektivität. S. 239258 in: C. Wolf und H. Best (Hrsg.): Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS.
Rhodes, G., 2006: The evolutionary psychology of facial beauty. Annual Review of Psychology 57: 199-226.
Rhodes, M. G. und J. S. Anastasi, 2012: The own-age bias in face recognition: A meta-analytic and theoretical review. Psychological Bulletin 138: 146-174.
Rosenthal, R., 1990: How are we doing in soft psychology? American Psychologist 45: 775–777.
Roszell, P., D. Kennedy und E. Grabb, 2001: Physical attractiveness and income attainment among Canadians. The Journal of Psychology 123: 547-559.
Rubenstein, A. J., J. H Langlois und L. A. Roggman, 2002: What makes a face attractive and why: The role of averageness in defining facial beauty. S. 1-33 in: G. Rhodes und L.A. Zebrowitz (Hrsg.): Facial attractiveness: Evolutionary, cognitive, and social perspectives. Ablex: Westport, CT.
Schuler, H. und W. Berger, 1979: Physische Attraktivität als Determinante von Beurteilung und Einstellungsempfehlung. Psychologie und Praxis 23: 59-70.
Sergl, H. G., 1991: Psychosoziale Auswirkungen einer Entstellung im Mund- und Gesichtsbereich. S. 19-27 in: H.G. Sergl und H. Müller-Fahlbusch (Hrsg.): Jahrbuch der Psychologie und der Psychosomatik in der Zahlheilkunde, Bd. 2. Berlin: Quintessenz.
Shackelford, T. K. und R. J. Larsen, 1999: Facial attractiveness and physical health. Evolution and Human Behavior 20: 71-76.
Shrout, P. E., 1998: Measurement reliability and agreement in psychiatry. Statistical Methods in Medical Research 7: 301-317.
Shrout, P. E. und J. L. Fleiss, 1979: Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 86: 420-428.
Sigall, H. und Ostrove, N., 1975: Beautiful but dangerous: Effects of offender attractiveness and nature of the crime on juridic judgment. Journal of Personality and Social Psychology 31: 410-414.
Singh, D., B. J. Dixson, T. S. Jessop, B. B. Morgan und A. F. Dixson, 2010: Cross-cultural consensus for waist–hip ratio and women's attractiveness. Evolution and Human Behavior 31: 176-181.
Teuscher, U. und C. Teuscher, 2007: Reconsidering the double standard of aging: Effects of gender and sexual orientation on facial attractiveness ratings. Personality and Individual Differences 42: 631-639.
Thornhill, R. und S. W. Gangestad, 1999: Facial attractiveness. Trends in Cognitive Sciences 3: 452-460.
Townsend, J. M. und G. D. Levy, 1990: Effects of potential partners' physical attractiveness and socioeconomic status on sexuality and partner selection. Archives of Sexual Behavior 19: 149-164.
Tramitz, C., 2000: Die Annäherung – der Erstkontakt zwischen Mann und Frau. S. 33-52 in: P. Kaiser (Hrsg.): Partnerschaft und Paartherapie. Göttingen: Hogrefe.

Umberson. D. und M. Hughes, 1987: The impact of physical attractiveness on achievement and psychological well-being. Social Psychology Quarterly 50: 227-236.
Zebrowitz, L. A., J. A. Hall, N. A. Murphy und G. Rhodes, 2002: Looking smart and looking good: Facial cues to intelligence and their origins. Personality and Social Psychology Bulletin 28: 238-249.

| | |
|---|---|
| Anschrift des Autors | Johannes Lutz<br>Universität Potsdam<br>Department Psychologie / Sozialpsychologie<br>Karl-Liebknecht-Str. 24-25<br>14476 Potsdam<br>jlutz@uni-potsdam.de |
| Ko-Autor/-innen | Christoph J. Kemper<br>Institut für medizinische und pharmazeutische Prüfungsfragen (IMPP), Mainz<br><br>Constanze Beierlein<br>GESIS – Leibniz-Institut für Sozialwissenschaften Mannheim<br><br>Jutta Margraf-Stiksrud<br>Philipps-Universität Marburg<br><br>Beatrice Rammstedt<br>GESIS – Leibniz-Institut für Sozialwissenschaften Mannheim |

methoden, daten, analysen · 2013, Jg. 7(2), S. 233-249 DOI: 10.12758/mda.2013.013

# Eine kurze Skala zur Messung der fünf Dimensionen der Persönlichkeit

## *10 Item Big Five Inventory (BFI-10)*

# A Short Scale for Assessing the Big Five Dimensions of Personality

## *10 Item Big Five Inventory (BFI-10)*

*Beatrice Rammstedt, Christoph J. Kemper, Mira Céline Klein, Constanze Beierlein, Anastassiya Kovaleva*

*Zusammenfassung*

Mit der zunehmenden Etablierung des Fünf-Faktoren-Modells der Persönlichkeit werden die sogenannten „Big Five"-Persönlichkeitsdimensionen vermehrt auch in Anwendungskontexten außerhalb der Psychologie und insbesondere auch in der Large-Scale-Forschung erhoben. Da jedoch gerade in diesen Bereichen die Untersuchungszeit oft stark begrenzt ist, sind die herkömmlichen Verfahren zur Erfassung der Fünf Faktoren in der Regel zu umfangreich. Speziell für solche Kontexte wurde das ultra-ökonomische BFI-10 entwickelt, das die fünf Dimensionen mit insgesamt zehn Fragen bzw. einer durchschnittlichen Bearbeitungsdauer von circa einer Minute erfasst. Das BFI-10 wurde in der vorliegenden Studie an einer umfangreichen, bevölkerungsrepräsentativen Stichprobe validiert. Die Ergebnisse belegen zufriedenstellende psychometrische Kennwerte für das BFI-10. Darüber hinaus konnten die Konstrukt- und die Kriteriumsvalidität des Verfahrens empirisch untermauert werden.

*Abstract*

In course of the further establishment of the Five-Factor Model as the predominant model for describing personality there is an increasing demand for assessing the Big Five dimensions of personality in contexts outside the core field of psychology. As in such contexts assessment time is severely limited, established Big Five instruments are generally too lengthy. Especially for such contexts we developed the ultra-short BFI-10 assessing the five dimensions with a total of ten items, respectively an average duration of approximately one minute. The present study validated the BFI-10 based on a large and population representative sample. Results indicate sufficient psychometric properties for the BFI-10 scales and items. In addition, the findings corroborate the construct and criterion validity of the instrument.

*Keywords*: BFI-10, Big Five, Kurzversion, Kurzskala, Persönlichkeit
*Schlüsselworte*: BFI-10, Big Five, short version, short scale, personality

# 1 Einleitung

Ein weit verbreitetes und akzeptiertes Modell der Persönlichkeit ist das so genannte Fünf-Faktoren-Modell (vgl. De Raad 2000; Goldberg 1990; John/Naumann/Soto 2008). Dieses Modell basiert auf einem lexikalischen Ansatz, der annimmt, dass alle wesentlichen interindividuellen Differenzen einer Sprache im Wörterbuch durch entsprechende Begriffe repräsentiert sind. In einer Vielzahl von Studien (Borgotta 1964; Digman/Takemoto-Chock 1981; Norman 1963; Tupes/Christal 1961) konnte belegt werden, dass sich die Einschätzung von Personen auf diesen persönlichkeitsbeschreibenden Begriffen auf globalster Ebene auf fünf Dimensionen – auch „Big Five" genannt (vgl. Goldberg 1981) – aggregieren lässt. Die erste Dimension, *Extraversion*, subsumiert Merkmale wie Geselligkeit, Aktivität, Gesprächigkeit und Durchsetzungsfähigkeit auf der einen Seite, während der Introversionspol durch Begriffe wie still, schweigsam und zurückgezogen charakterisiert ist. Die zweite Dimension, *Verträglichkeit*, beschreibt interpersonelles Verhalten. Personen mit einer hohen Ausprägung auf diesem Faktor sind altruistisch, neigen zu zwischenmenschlichem Vertrauen, zur Kooperativität und zur Nachgiebigkeit. Personen mit einer niedrigen Ausprägung in der Dimension *Verträglichkeit* lassen sich als kühl, kritisch und misstrauisch beschreiben. Die Dimension *Gewissenhaftigkeit* differenziert Personen, die zielstrebig, ausdauernd, diszipliniert und zuverlässig sind, von solchen, die nachlässig, gleichgültig und unbeständig sind. Die vierte Dimension, *Neurotizismus*, beschreibt, wie emotional labil eine Person reagiert. Personen mit hoher Ausprägung auf diesem Faktor neigen zu Unsicherheit, reagieren eher nervös, ängstlich und deprimiert. Der fünfte Faktor, *Offenheit für Erfahrungen,* umfasst Aspekte wie das Interesse an neuen Erfahrungen, Erlebnissen und Eindrücken. Personen mit einer hohen Ausprägung auf diesem Faktor sind wissbegierig, phantasievoll, intellektuell und künstlerisch interessiert, während Personen mit einer niedrigen Ausprägung eher zu festen Ansichten, wenig Interesse an Neuem und Konservatismus neigen.

Auch wenn die Big Five-Dimensionen der Persönlichkeit erstmals für den nordamerikanischen Sprachraum identifiziert wurden, belegen zahlreiche Studien basierend auf separaten, landesspezifischen lexikalischen Ansätzen ihre Angemessenheit auch für viele andere Sprachräume. Als eine der besten Replikationen erwies sich hierbei die deutsche Taxonomie (Ostendorf 1990).

Inzwischen ist dieses Persönlichkeitsmodell auch außerhalb der persönlichkeitspsychologischen Forschung stark verbreitet. So haben sich die Big Five-Persönlichkeitsdimensionen in den letzten Jahren als erfolgreiche Prädiktoren für verschiedene individuelle wie gesellschaftliche Prozesse und Phänomene erwiesen (für einen Überblick s. Ozer/Benet-Martinez 2006). Beispielsweise konnte gezeigt werden, dass emotional stabilere und gewissenhaftere Personen in der Regel gesünder und länger leben (vgl. Roberts/Kuncel/Shiner/Caspi/Goldberg 2007). Auch die Berufswahl und der berufliche Erfolg werden von der individuellen Persönlichkeit geprägt. So bevorzugen verträglichere Personen soziale Berufe, gewissenhaftere eher konventionelle Tätigkeiten, z.B. im Bereich Verwaltung und Management, und offene Personen forschende oder künstlerische Tätigkeiten (vgl. Barrick/Mount/Gupta 2003). Gewissenhaftigkeit ist neben der allgemeinen kognitiven Leistungsfähigkeit der zentrale Prädiktor für den beruflichen Erfolg (Schmidt/Hunter 1998). Die Zufriedenheit mit der Arbeit ist tendenziell höher bei extravertierten, emotional stabileren und gewissenhafteren Personen (Judge/Heller/Mount 2002). Auch für das Wahlverhalten ließen sich Zusammenhänge mit der Persönlichkeit nachweisen. So wählen Gewissenhafte eher Parteien, die im Parteienspektrum mitte-rechts angesiedelt sind, während offene und verträgliche Menschen eher mitte-links wählen (Vecchione/Schoen/González-Castro/Cieciuch/Pavlopoulos/Caprara 2011).

Die Ergebnisse dieser Studien lassen es sinnvoll erscheinen, Persönlichkeit auch in Studien zu erfassen, in denen die Persönlichkeit zwar nicht von primärem Interesse ist, aber eine möglicherweise sehr interessante und nützliche Variable darstellen kann. Zu denken ist hier beispielsweise an sozial- oder wirtschaftswissenschaftliche Large-Scale-Studien wie das Sozio-ökonomische Panel des Deutschen Instituts für Wirtschaftsforschung. Da in solchen Forschungskontexten neben Persönlichkeit meist auch zahlreiche weitere Merkmale erhoben werden, ist die Zeit, die hier zur Erfassung der Persönlichkeitsvariablen aufgewendet werden kann, in der Regel stark begrenzt. Die etablierten Verfahren zur Erfassung der Big Five stammen jedoch zumeist aus dem Bereich der persönlichkeitspsychologischen Individualdiagnostik und sind für diese Zwecke entsprechend umfangreich konzipiert, um eine differenzierte Erfassung der Konstrukte zu ermöglichen. Für die hier beschriebenen Anwendungen außerhalb des persönlichkeitspsychologischen Feldes sind sie in der Regel jedoch zu zeitaufwändig. So umfasst das wohl bekannteste Inventar zur Messung der Big Five, das NEO-Personality Inventory, in seiner revidierten Form (NEO-PI-R, Costa/McCrae 1992; deutsche Adaptation: Ostendorf/Angleitner 2004) 240 Items, und selbst dessen Kurzform, das NEO-Five Factor Inventory (NEO-FFI; Costa/McCrae 1992; deutsche Adaptation: Borkenau/Ostendorf 1993) immerhin noch 60 Items. Eines der kürzesten standardisierten Ver-

fahren zur Erfassung der Big Five ist das 44 Items umfassende Big Five Inventory (BFI; John/Donahue/Kentle 1991; psychometrische Kennwerte werden beschrieben in Benet-Martinèz/John 1998 und in John/Naumann/Soto 2008; deutsche Adaptation: Rammstedt 1997; s.a. Lang/Lüdtke/Asendorpf 2001). Das BFI wurde mit dem Ziel entwickelt, die prototypischen fünf Faktoren zu erfassen und somit den gemeinsamen Kern der verschiedenen Fünf-Faktoren-Ansätze abzubilden (für Details vgl. John/Srivastava 1999). Validierungsstudien zeigen, dass das BFI neben seiner relativen Ökonomie ein sehr reliables und valides Instrument zur Erfassung der Big Five darstellt (Benet-Martinèz/John 1998; John/Srivastava 1999). Ähnliche Ergebnisse ließen sich auch für die deutschen Adaptationen erzielen (Lang et al., 2001; Rammstedt 1997). Auch wenn das BFI somit bereits ein vergleichsweise ökonomisches Inventar zur Erfassung der Big Five darstellt, ist dessen Bearbeitungszeit von 5 bis 10 Minuten für viele der oben angeführten Untersuchungskontexte zu lang. Basierend auf dem BFI wurden daher im Laufe der letzten Jahre verschiedene Kurzformen entwickelt, wie das 21 Items umfassende BFI-K (Rammstedt/John 2005), das 15 Items umfassende BFI-S (Schupp/Gerlitz 2008) und das 10 Items umfassende BFI-10 (Rammstedt/John 2007). Aufgrund seiner hohen Ökonomie und seiner für Gruppenuntersuchungen adäquaten psychometrischen Güte werden insbesondere das BFI-10 sowie das BFI-S in zahlreichen sozial- und wirtschaftswissenschaftlichen Large-Scale-Studien wie dem Sozio-ökonomischen Panel oder dem International Social Survey Programme (ISSP) eingesetzt.

Das BFI-10 wurde ursprünglich basierend auf rein studentischen Stichproben entwickelt und validiert (Rammstedt/John 2007) und im Anschluss basierend auf einer bevölkerungsrepräsentativen Stichprobe normiert (Rammstedt 2007a). Ziel der vorliegenden Studie ist es zunächst, die Güte des BFI-10 basierend auf einer aktuellen und heterogenen bevölkerungsrepräsentativen Stichprobe zu replizieren und somit einen weiteren Beleg für die Konstruktvalidität der Kurzskala zu präsentieren. Des Weiteren soll der vorliegende Artikel potentiellen Nutzerinnen und Nutzern, insbesondere auch aus dem Bereich der Large-Scale-Forschung, ein für zeitlich stark limitierte Erhebungskontexte angemessenes Erhebungsinstrument zur Erfassung der fünf Hauptdimensionen der Persönlichkeit zur Verfügung stellen. Damit ist die Hoffnung verknüpft, dass durch den vermehrten Einsatz dieses standardisierten psychologischen Erhebungsinstruments eine erhöhte Anschlussfähigkeit und Vergleichbarkeit zwischen Untersuchungen und eine verbesserte Deskription und Prädiktion wissenschaftlich und gesellschaftlich relevanter Prozesse und Phänomene erzielt werden kann.

# 2 Methode

## 2.1 Stichprobe

Die Hauptstichprobe bestand aus *N* = 1.134 Personen, die repräsentativ für die Wohnbevölkerung in Deutschland mit einem Alter von über 17 Jahren sind. Sie wurde mithilfe des ADM-Stichprobensystems der Arbeitsgemeinschaft deutscher Marktforschungsinstitute zufällig gezogen. Die gesamte Erhebung (siehe Abschnitt 2.2) erfolgte in Form eines computergestützten Interviews (mittlere Dauer: *M* = 43 Minuten, *SD* = 13 Minuten) und wurde von einem unabhängigen kommerziellen Anbieter durchgeführt.

Ergänzend hierzu wurde an einer quotierten Stichprobe (geschichtet nach den Merkmalen Geschlecht, Alter, Bildung und Bundesland) die zeitliche Stabilität der Persönlichkeitseinschätzungen untersucht. Die Erhebung dieser Retest-Stichprobe (*N* = 338) erfolgte in zwei Wellen mit einem zeitlichen Abstand von 6 bis 10 Wochen. In Welle 1 nahmen *N* = 239 Befragte im CAPI-Modus und *N* = 99 im Selbstausfüller-Modus teil. Zu Welle 2 wurden *N* = 227 Personen im CAPI-Modus und *N* = 111 im Selbstausfüller-Modus befragt. Dabei wurden drei Bedingungen realisiert, zu denen die Teilnehmer/-innen randomisiert zugeordnet wurden: *N* = 128 Befragte beantworten die Items zu beiden Wellen im CAPI-Modus, *N* = 111 Befragte erhielten in Welle 1 die CAPI-Version und in Welle 2 die Selbstausfüllerversion, *N* = 99 Befragte beantworten erst die Selbstausfüller- und dann in Welle 2 die CAPI-Version.[1] Die gesamte Befragung dauerte im Durchschnitt *M* = 53 Minuten (*SD* = 12 Minuten). Die Charakteristika der Haupt- und der Retest-Stichprobe sind in Tabelle 1 dargestellt (für weitere Details siehe Kemper/Beierlein/Kovaleva/Rammstedt in Druck).

1 Auf eine Auswertung der Substichprobe, die zu beiden Messzeitpunkten das BFI-10 im Selbstausfüller-Modus bearbeitete, wurde aufgrund des geringen Stichprobenumfangs verzichtet.

Tabelle 1 Charakteristika der Stichproben

| | | Hauptstichprobe | Retest-Stichprobe |
|---|---|---|---|
| *Stichprobe* | | | |
| Umfang [*N*] | | 1.134 | 338 |
| Art | | Zufall | Quote |
| Modus | | CAPI | CAPI, Papier |
| *Zusammensetzung* | | | |
| Geschlecht [% m.] | | 44,4% | 47,9% |
| Alter [*M* (*SD*)] | | 53.3 (18.4) | 46.7 (15.1) |
| Bildung | ≤ 9 Jahre | 37,2% | 45,3% |
| | 10 Jahre | 37,0% | 27,9% |
| | ≥ 11 Jahre | 25,8% | 25,4% |

*Anmerkungen: CAPI = Computer Assisted Personal Interview, Papier = Papierversion (Selbstausfüller).*

## 2.2 Instrumente

### *Das BFI-10*

Das BFI-10 (Rammstedt/John 2007; Rammstedt 2007a) ist eine Kurzversion des etablierten Big Five Inventory (BFI; John et al. 1991). Es erfasst die Big Five-Persönlichkeitsdimensionen mit zwei Items pro Dimension, von denen jeweils eines den positiven und eines den negativen Pol abbildet. Ferner wurde im Zuge der Itemselektion intendiert, mit den beiden selektierten Items eine möglichst maximale Bandbreite der entsprechenden Dimension abzubilden, um zu vermeiden, dass beide Items nur eine Facette der in Frage stehenden Dimension abdecken. Hiermit wurde jedoch in Kauf genommen, dass die beiden Items einer Dimension nur mittelmäßige Korrelationen untereinander und daher nur geringe interne Konsistenzen aufweisen. Die Items sind mittels einer fünfstufigen Ratingskala von „trifft überhaupt nicht zu" (1) bis „trifft voll und ganz zu" (5) zu beantworten. Die Administrierung des BFI-10 im computerbasierten Modus (CAPI) dauert etwa 80 Sekunden (Median). Die Items sind in Appendix A dargestellt (detailliertere Informationen zur Skala mit Auswertungshinweisen und englischer Fragebogenversion finden sich bei Rammstedt/Kemper/Klein/Beierlein/Kovaleva 2012). Rammstedt und John (2007) überprüften die psychometrische Güte ihrer auf diese Weise entwickelten Kurzskala auf Basis von vier studentischen Stichproben und konnten zufriedenstellende Reliabilitäts- und Validitätskennwerte nachweisen.

Tabelle 2 Deskriptive Statistiken des BFI-10 für die Hauptstichprobe

| | *M* | *SD* | Sch | Kurt |
|---|---|---|---|---|
| Ich bin eher zurückhaltend, reserviert. | 2.88 | 1.22 | -0.02 | -1.15 |
| Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen. | 3.48 | 1.02 | -0.51 | -0.37 |
| Ich bin bequem, neige zur Faulheit. | 2.00 | 1.10 | 0.90 | -0.12 |
| Ich bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen. | 3.40 | 1.07 | -0.48 | -0.63 |
| Ich habe nur wenig künstlerisches Interesse. | 2.96 | 1.31 | -0.00 | -1.29 |
| Ich gehe aus mir heraus, bin gesellig. | 3.83 | 1.04 | -0.74 | -1.19 |
| Ich neige dazu, andere zu kritisieren. | 2.58 | 1.09 | 0.27 | -0.76 |
| Ich erledige Aufgaben gründlich. | 4.30 | 0.83 | -1.31 | 1.81 |
| Ich werde leicht nervös und unsicher. | 2.24 | 1.07 | 0.68 | -0.25 |
| Ich habe eine aktive Vorstellungskraft, bin fantasievoll. | 3.77 | 1.03 | -0.71 | -0.00 |

*Anmerkungen: Sch = Schiefe, Kurt = Kurtosis.*

Die Fragebogenbatterien der Retest- und der Hauptstichprobe beinhalteten neben dem BFI-10 umfangreiche soziodemographische Angaben sowie weitere psychologische und sozialwissenschaftliche Validierungsmaße. Dabei wurden sowohl etablierte Messinstrumente der sozialwissenschaftlichen Forschung eingesetzt (z.B. Fragen zu verschiedenen Aspekten der Lebenszufriedenheit aus SOEP 2005 und 2009 und Fragen zum selbstberichteten devianten Verhalten aus dem ALLBUS 2010) als auch psychologische Skalen (z.B. zur Erfassung von Lebenszufriedenheit: SWLS, Diener/Emmons/Larsen/Griffin 1985; Kontrollüberzeugungen: Jakoby/Jacob 1999) als auch eigens entwickelte Kurzskalen (z.B. interpersonelles Vertrauen mittels KUSIV3: Beierlein/Kemper/Kovaleva/Rammstedt 2012). Die Items zu den soziodemographischen Angaben wurden größtenteils den demographischen Standards des Statistischen Bundesamtes (2010) entnommen.

## 3 Ergebnisse

Tabelle 2 zeigt die Mittelwerte und Standardabweichungen, Kurtosis und Schiefe der zehn BFI-10-Items. Die Items waren größtenteils normalverteilt und wiesen keine Decken- oder Bodeneffekte auf. Lediglich das Item „Ich erledige Aufgaben gründlich." war im Mittelwert erhöht und auffällig linksschief.

Tabelle 3 Reliabilitätskoeffizienten (Retest) des BFI-10

| Big-Five-Dimension | Reteststichprobe (gesamt; *N* = 338) | Reteststichprobe (CAPI/CAPI; *N* = 128) | Reteststichprobe (CAPI/Selbstausfüller; *N* = 111) | Reteststichprobe (Selbstausfüller/CAPI; *N* = 99) | Rammstedt & John (2007)[1] |
|---|---|---|---|---|---|
| Extraversion | .59 | .67 | .59 | .44 | .84 |
| Verträglichkeit | .50 | .49 | .53 | .47 | .58 |
| Gewissenhaftigkeit | .59 | .63 | .63 | .53 | .77 |
| Neurotizismus | .49 | .56 | .46 | .41 | .74 |
| Offenheit | .62 | .67 | .64 | .54 | .72 |
| Mittelwert | .56 | .60 | .57 | .48 | .73 |

*Anmerkungen: 1 Stichprobe 1 (G-1), N = 184, Retest-Intervall: 6 Wochen.*

Aufgrund der extrem geringen Itemanzahl pro Skala und der intendierten Heterogenität der beiden Items einer Skala eignete sich die interne Konsistenz nicht als guter Schätzer für die Reliabilität der fünf Skalen. Stattdessen wurde die Retest-Reliabilität über ein Intervall von sechs bis acht Wochen bestimmt. In dem Retestdesign wurde teilweise der Erhebungsmodus von CAPI zu Paper-Pencil variiert. Um eine Konfundierung der Retestreliabilität mit Effekten des Erhebungsmodus zu vermeiden bzw. diese abschätzen zu können, sind in Tabelle 3 die Retestkoeffizienten sowohl für die gesamte Reteststichprobe als auch separat für die verschiedenen Teilstichproben, je nach Kombination des Erhebungsmodus zu den beiden Messzeitpunkten, dargestellt. Zum Vergleich sind in Tabelle 3 ebenfalls die von Rammstedt und John basierend auf ihrer ersten Validierungsstudie (2007) berichteten Retest-Reliabilitätskoeffizienten aufgeführt. Im Mittel liegt die Retest-Reliabilität für die Gesamtstichprobe bei $r_{tt}$= .56. Unter Konstanthaltung des Erhebungsmodus (Teilstichprobe CAPI/CAPI) resultiert eine leicht höhere Stabilität von $r_{tt}$= .60 im Mittel. Bei Variation des Erhebungsmodus zeigen sich wie zu erwarten leicht niedrigere Retest-Koeffizienten als unter Konstanthaltung ($r_{tt}$= .57 und $r_{tt}$= .48). Im Vergleich zu der ursprünglichen Validierungsstudie basierend auf rein studentischen Daten mit einer mittleren Stabilität von $r_{tt}$= .73 sind demnach die hier gefundenen Retest-Reliabilitäten mit einem Mittel von $r_{tt}$= .56 (Gesamtstichprobe) bzw. .60 (bei konstantem Erhebungsmodus) leicht niedriger. Sowohl in der Gesamtstichprobe als auch in der erhebungsmoduskonstanten Stichprobe finden sich die höchsten Koeffizienten für Offenheit ($r_{tt}$= .62/.67). Verträglichkeit ($r_{tt}$= .50/.49) weist hingegen eine vergleichsweise niedrige Retest-Reliabilität bzw. Stabilität auf; ein Ergebnis, das häufig in der Literatur zur Stabilität von Big

Tabelle 4 Varimax rotierte faktorielle Struktur der BFI-10 Items für die Hauptstichprobe

| Item | E | V | G | N | O |
|---|---|---|---|---|---|
| Ich bin eher zurückhaltend, reserviert. | **-.77** | -.02 | .11 | .12 | .09 |
| Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen. | .27 | **-.81** | .12 | .16 | .02 |
| Ich bin bequem, neige zur Faulheit. | .00 | .16 | **-.77** | .07 | -.02 |
| Ich bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen. | .03 | -.01 | .05 | **-.79** | .01 |
| Ich habe nur wenig künstlerisches Interesse. | -.01 | .07 | .02 | -.05 | **.92** |
| Ich gehe aus mir heraus, bin gesellig. | **.77** | -.11 | .20 | -.12 | -.04 |
| Ich neige dazu, andere zu kritisieren. | .34 | **.66** | -.02 | .35 | .08 |
| Ich erledige Aufgaben gründlich. | .09 | .04 | **.82** | -.08 | -.09 |
| Ich werde leicht nervös und unsicher. | -.25 | .01 | -.12 | **.76** | .03 |
| Ich habe eine aktive Vorstellungskraft, bin fantasievoll. | .45 | .09 | .21 | -.14 | **-.54** |

*Anmerkungen: E= Extraversion, V = Verträglichkeit, G = Gewissenhaftigkeit, N = Neurotizismus, O = Offenheit; höchste Itemladungen sind fett gedruckt.*

Five-Inventaren berichtet wird (z.B. Caruso 2000; Hahn/Gottschling/Spinath 2012; Muck/Hell/Gosling 2007; Romero/Villar/Gómez-Fraguela/López-Romero 2012). Zur Bestimmung der faktoriellen Validität des BFI-10 wurde geprüft, ob sich die angenommene fünffaktorielle Struktur des Big Five-Ansatzes mit den BFI-10-Items replizieren lässt. Hierzu wurden die Daten der Hauptstichprobe einer Hauptkomponentenanalyse mit anschließender Varimax-Rotation unterzogen. Gemäß dem Fünf-Faktoren-Modell wurden fünf Faktoren extrahiert. Wie aus der in Tabelle 4 dargestellten Ladungsmatrix erkennbar, weisen alle Items hypothesenkonform ihre höchste Ladung auf dem korrespondierenden Faktor auf. Außerdem weisen nur zwei Items („Ich neige dazu, andere zu kritisieren." und „Ich habe eine aktive Vorstellungskraft, bin fantasievoll.") Sekundärladungen größer .30 auf. Demnach spiegelt die faktorielle Struktur des BFI-10 sehr gut die intendierte fünf-faktorielle Einfachstruktur wider.

Rammstedt und John (2007) konnten zeigen, dass das BFI-10 hohe Übereinstimmung nicht nur mit der BFI-Gesamtskala, sondern auch mit dem NEO-PI-R und dessen Facetten aufweist. Im Rahmen der vorliegenden Studie sollte darüber hinaus geprüft werden, ob sich mit dem BFI-10 die in der Literatur typischerweise berichteten Korrelationen zwischen Persönlichkeitsmerkmalen und soziodemogra-

Tabelle 5 Validitätskoeffizienten des BFI-10 für die Hauptstichprobe

| | E | V | G | N | O |
|---|---|---|---|---|---|
| *Soziodemographische Variablen* | | | | | |
| Alter | -.21** | .08** | .17** | -.03 | -.10** |
| Geschlecht[1] | .04 | .06* | .12** | .21** | .10** |
| Einkommen[2] | .12** | -.05 | .04 | -.17** | .07 |
| Bildung: Schuljahre[2] | .16** | -.04 | -.01 | -.05 | .24** |
| Bildung: Bücher[2] | .14** | .02 | -.00 | -.06* | .25** |
| *Zufriedenheit* | | | | | |
| Leben | .28** | .14** | .25** | -.29** | .19** |
| Arbeit | .13** | .13** | .29** | -.20** | .14** |
| Partner | .17** | .07 | .23** | -.18** | .12** |
| Körperliche Gesundheit | .22** | .04 | .15** | -.25** | .16** |
| Psychische Gesundheit | .23** | .06* | .19** | -.35** | .09** |
| *Devianz[2]* | | | | | |
| Fahren ohne Fahrkarte | .08** | -.10** | -.24** | .07* | .06* |
| Autofahren mit über 0.5 Promille | .07* | -.14** | -.14** | -.03 | -.04 |
| Diebstahl | .02 | -.11** | -.22** | .05 | .01 |
| Steuerhinterziehung | .07* | -.09** | -.11** | -.05 | -.05 |
| Größe soziales Netz[3] | .12** | -.02 | .03 | .04 | .14** |
| Interpersonelles Vertrauen | .16** | .35** | .14** | -.16** | 16.** |

*Anmerkungen: E= Extraversion, V = Verträglichkeit, G = Gewissenhaftigkeit, N = Neurotizismus, O = Offenheit; [1] männlich=0, weiblich=1; [2] Rangkorrelation nach Spearman; [3] Anzahl der genannten Personen auf die Frage: „Hin und wieder besprechen die meisten Leute wichtige Angelegenheiten mit Anderen: Wenn Sie an die letzten sechs Monate zurückdenken, mit wem haben Sie über Dinge gesprochen, die Ihnen wichtig waren?"; * p < .05, ** p < .01.*

phischen sowie verschiedenen sozialwissenschaftlichen Inhaltsvariablen replizieren lassen. Tabelle 5 zeigt die berichteten Zusammenhänge. Insgesamt betrachtet lassen sich in Bezug auf die soziodemographischen Variablen Alter und Geschlecht die bereits etablierten Zusammenhänge weitgehend replizieren: Extraversion korreliert negativ und Gewissenhaftigkeit positiv mit dem Alter (vgl. Caspi/Roberts/Shiner 2005; Lucas/Donnellan 2011; McCrae et al. 1999). In Bezug auf Geschlechtsunterschiede zeigt sich auch für das BFI-10 der bekannte Effekt, dass Frauen im Vergleich zu Männern signifikant höhere Werte in Neurotizismus aufweisen (vgl. Costa/Terracciano/McCrae 2001; Feingold 1994; Körner/Geyer/Brähler 2002; Srivastava/John/Gosling/Potter 2002; Viken/Rose/Kapiro/Koskenvuo 1994). Für Bildung, gemessen über die Anzahl der Schuljahre sowie kulturelles Kapital, gemessen

über die Anzahl der Bücher im Elternhaus, lässt sich der bekannte Effekt replizieren, dass gebildetere Personen eine höhere Offenheit für Erfahrungen aufweisen (vgl. Caspi et al. 2005; Goldberg/Sweeney/Merenda/Hughes 1998; Körner et al. 2002; Vassend/Skrondal 1995).

In Bezug auf die untersuchten Inhaltsvariablen zeigen die BFI-10-Skalen geringe bis mittlere Korrelationen (vgl. Cohen 1992) mit verschiedenen globalen und spezifischen Maßen der Lebenszufriedenheit. Neurotizismus korreliert erwartungskonform durchweg negativ mit den Zufriedenheitsmaßen, wie vergleichbar auch schon von Rammstedt (2007b) oder in Bezug auf Ehezufriedenheit von Kelly und Conley (1987) berichtet wurde. Judge, Heller und Mount (2002) konnten in ihrer Metaanalyse positive Zusammenhänge zwischen Arbeitszufriedenheit und Extraversion, Verträglichkeit und Gewissenhaftigkeit sowie einen negativen Zusammenhang zwischen Arbeitszufriedenheit und Neurotizismus nachweisen. Diese Befunde decken sich mit den hier vorliegenden Korrelationsmustern.

Die Größe des sozialen Netzwerks, erfasst über die Anzahl der Personen, mit denen eine Befragungsperson wichtige Angelegenheiten bespricht, sollte positiv mit Extraversion und Offenheit assoziiert sein, da Extravertierte die Gesellschaft anderer Menschen suchen und offene Personen neue Erfahrungen und Situationen aufsuchen, in denen soziale Kontakte entstehen können (vgl. Amelang/Bartussek 2001; Gosling/Augustine/Vazire/Holtman/Gaddis 2011; Selfhout et al. 2010). Diese Annahmen werden durch die Ergebnisse bestätigt. Deviantes Verhalten wie Autofahren ohne Führerschein oder Diebstahl sollte insbesondere einen negativen Zusammenhang mit Gewissenhaftigkeit aufweisen (Roberts/Walton/Bogg/Caspi 2006; Walton/Roberts 2004). Auch diese Annahme bestätigt sich für alle vier untersuchten Indikatoren devianten Verhaltens. Interpersonelles Vertrauen schließlich spiegelt einen Aspekt der Big Five-Dimension Verträglichkeit wider (vgl. Amelang/Bartussek 2001; Ostendorf/Angleitner 2004; Sneed 2002) und sollte daher primär mit dieser Dimension in Zusammenhang stehen. Auch diese Annahme wird durch die Befunde gestützt.

## 4 Diskussion

Ziel der vorliegenden Untersuchung war es, das BFI-10 an einer bevölkerungsrepräsentativen Stichprobe zu validieren und somit für Untersuchungskontexte mit starken Zeitlimitationen eine reliable und valide Erfassung der Fünf Faktoren zu ermöglichen. Das BFI-10 wurde hierzu mittels einer umfangreichen, bevölkerungsrepräsentativen Stichprobe auf seine psychometrischen Kennwerte hin untersucht.

Mit einer Ausnahme konnte für alle Items eine mittlere Schwierigkeit und Normalverteilung nachgewiesen werden. Die fünf Skalen des BFI-10 erwiesen sich als hinreichend reliabel und stabil. Im Vergleich zu den von Rammstedt und John (2007) berichteten Stabilitätskennwerten fallen für jede Dimension – auch innerhalb der CAPI-Stichprobe – die in der aktuellen Studie ermittelten geringer aus. Dies mag mehrere Ursachen haben. So unterscheiden sich die Studien sowohl im Design, nämlich im Retestintervall, welches in der vorliegenden Studie etwas länger war, als auch im Erhebungsmodus und in der untersuchten Stichprobe. Die hier wiederholt erhobene Stichprobe war deutlich bildungs- und altersheterogener und vermutlich weniger geübt im Umgang mit Befragungen als die von Rammstedt und John untersuchten studentischen Stichproben. In der Studie von Rammstedt und John wurde der Fragebogen als Teil einer Fragebogenbatterie schriftlich vorgegeben (PAPI). In der vorliegenden Studie war das BFI-10 zwar auch eingebettet in eine Fragebogenbatterie, diese wurde jedoch als Interview erhoben. Gemessene Retest-Stabilitäten sind insbesondere bei kürzeren Messwiederholungsintervallen zumeist konfundiert mit Erinnerungseffekten. Es lässt sich vermuten, dass diese Erinnerungseffekte im Interview (ohne visuelle Informationen über die eigene Antwort) geringer ausfallen als im PAPI-Modus, in dem die Befragungsperson die Items einsehen und beantworten kann. Dies mag eine mögliche Erklärung für die hier gefundenen geringeren Stabilitätskoeffizienten sein. Aufgrund des Designs der vorliegenden Studie kann sie hier jedoch nicht geprüft werden und müsste daher in zukünftigen Studien näher beleuchtet werden.

Darüber hinaus erwies sich das BFI-10 in der vorliegenden Studie als hinreichend valide, sowohl in Hinblick auf die faktorielle Struktur der zehn Items als auch in Hinblick auf die hypothesenkonformen und in der Fachliteratur wiederholt berichteten Korrelationen mit soziodemographischen und sozio-ökonomischen Merkmalen.

Zukünftige Studien sollten jedoch noch näher die psychometrischen Kennwerte inklusive der prädiktiven Validitätskoeffizienten des BFI-10 direkt mit den etablierten Langversionen des BFI oder des NEO-PI-R vergleichen. Dieser Vergleich würde eine erste Abschätzung des durch die Verwendung einer solchen ultrakurzen Big Five-Version in Kauf genommenen Reliabilitäts- und Validitätsverlusts ermöglichen.

Auch wäre es wünschenswert, dass in zukünftige Studien zusätzliche Maße wie z.B. Fremdurteile in die Validierung mit einbezogen würden.

Zusammengefasst lässt sich konstatieren, dass mit dem BFI-10 ein extrem ökonomisches Instrument zur reliablen und validen Erfassung der Big Five-Persönlichkeitsdimensionen zur Verfügung steht. In der vorliegenden Studie konnte

ferner nachgewiesen werden, dass sich das BFI-10 mit vergleichbarer Güte auch für eine Datenerhebung im Interviewmodus eignet. Somit bietet sich das BFI-10 insbesondere für Forschungskontexte an, in denen die Erhebungsdauer ein kritischer Kostenfaktor ist, wie beispielsweise sozialwissenschaftliche Umfragen.

## Referenzen

ALLBUS (Die Allgemeine Bevölkerungsumfrage der Sozialwissenschaften) (2010). [Datensatz und Codebook].

Amelang, M. und D. Bartussek, 2001: Differentielle Psychologie und Persönlichkeitsforschung. Stuttgart: Kohlhammer.

Barrick, M. R., M. K. Mount und R. Gupta, 2003: Meta-analysis of the relationship between the five-factor model of personality and Holland's occupational types. Personnel Psychology 56: 45-74.

Beierlein, C., C. J. Kemper, A. Kovaleva und B. Rammstedt, 2012: Kurzskala zur Messung des zwischenmenschlichen Vertrauens: Die Kurzskala Interpersonales Vertrauen (KUSIV3). Mannheim: Gesis.

Benet-Martínez, V. und O. John, 1998: Los Cinco Grandes across cultures and ethnic groups: Multitrait-Multimethod analyses of the Big Five in Spanish and English. Journal of Personality and Social Psychology 75: 729-750.

Borgotta, E. F., 1964. The structure of personality characteristics. Behavioral Science 9: 8-17.

Borkenau, P. und F. Ostendorf, 1993: NEO-Fünf-Faktoren Inventar (NEO-FFI) nach Costa und McCrae. Göttingen: Hogrefe.

Caruso, J. C., 2000: Reliability generalization of the NEO personality scales. Educational and Psychological Measurement 60: 236-254.

Caspi, A., B. W. Roberts und R. L. Shiner, 2005: Personality development: Stability and change. Annual Review of Psychology 56: 453-484.

Cohen, J., 1992: A power primer. Psychological Bulletin 112: 155–159.

Costa, P. T. und R. R. McCrae, 1992: Revised NEO Personality Inventory and NEO Five Factor Professional Manual. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., A. Terraciano und R. R. McCrae, 2001: Gender differences in personality traits across cultures: Robust and surprising findings. Journal of Personality and Social Psychology 81: 322-331.

De Raad, B., 2000: The Big Five personality factors. Seattle, WA: Hogrefe.

Diener, E. D., R. A. Emmons, R. J. Larsen und S. Griffin, 1985: The Satisfaction with life scale. Journal of Personality Assessment 49: 71-75.

Digman, J. und N. K. Takemoto-Chock, 1981. Factors in the natural language of personality: Reanalysis and comparison of six major studies. Multivariate Behavioral Research 18: 149-170.

Feingold, A., 1994: Gender differences in personality – a metaanalysis. Psychological Bulletin 116: 429-456.

Goldberg, L. R., 1981: Language and individual differences: The search for universals in personality lexicons. S. 141-165 in: L. Wheeler (Hg.): Review of Personality and Social Psychology, Beverly Hills: Sage.

Goldberg, L. R., 1990: An alternative "description of personality": The Big-Five factor structure. Journal of Personality and Social Psychology 59: 1216-1229.

Goldberg, L. R., D. Sweeney, P. F. Merenda und J. E. Hughes, 1998: Demographic variables and personality: The effects of gender, age, education, and ethnic/racial status on self-descriptions of personality attributes. Personality and Individual Differences 24: 393-403.

Gosling, S. D., A. A. Augustine, S. Vazire, N. Holtman und S. Gaddis, 2011: Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information. Cyberpsychology, Behavior, and Social Networking 14: 483-488.

Hahn, E., J. Gottschling und F. M. Spinath, 2012: Short measurement of personality - Validity and reliability of the GSOEP Big Five Inventory (BFI-S). Journal of Research in Personality 46: 355-359.

Jakoby, N. und R. Jacob, 1999: Messung von internen und externen Kontrollüberzeugungen. ZUMA-Nachrichten, 45: 61-71.

John, O. P. und S. Srivastava, 1999: The Big Five trait taxonomy: History, measurement, and theoretical perspectives. S. 102-138 in: L. A. Pervin und O. P. John (Hg.): Handbook of personality: Theory and research. New York: Guilford Press.

John, O. P., E. M. Donahue und R. L. Kentle, 1991: The Big Five Inventory – versions 4a and 5. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

John, O. P., L. P. Naumann und C. J. Soto, 2008: Paradigm shift to the integrative Big-Five trait taxonomy: History, measurement, and conceptual issues. S.114-158 in: O. P. John, R. W. Robins und L. A. Pervin (Hg): Handbook of personality: Theory and research. New York: Guilford Press.

Judge, T. A., D. Heller und M. K. Mount, 2002: Five-factor model of personality and job satisfaction: A meta-analysis. Journal of Applied Psychology 87: 530-541.

Kelly, E. L. und J. J. Conley, 1987: Personality and compatibility: A prospective analysis of marital stability and marital satisfaction. Journal of Personality and Social Psychology 52: 27-40.

Kemper, C. J., C. Beierlein, A. Kovaleva und B. Rammstedt, 2013: Entwicklung und Validierung einer ultrakurzen Operationalisierung des Konstrukts Optimismus-Pessimismus – Die Skala Optimismus-Pessimismus-2 (SOP2). Diagnostica 59: 119-129.

Körner, A., M. Geyer und E. Brähler, 2002: Das NEO-Fünf-Faktoren Inventar (NEO-FFI). Valididerung anhand einer deutschen Bevölkerungsstichprobe. Diagnostica 48: 19-27.

Lang, F. R., O. Lüdtke und J. B. Asendorpf, 2001: Validity and psychometric equivalence of the German version of the Big Five Inventory in young, middle-aged and old adults. Diagnostica 47: 111-121.

Lucas, R. E. und M. B. Donnellan, 2011: Personality development across the life span: longitudinal analyses with a national sample from Germany. Journal of Personality and Social Psychology 101: 847-861.

Norman, W. T., 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. Journal of Abnormal and Social Psychology 66: 574-583.

McCrae, R. R., P. T. Costa, M. Pedroso de Lima, A. Simoes, F. A. Angleitner, I. Marusic, D. Bratko, G. V. Caprara, C. Barbaranelli, J.-H. Chae und R. L. Piedmont, 1999: Age differences in personality across the adult life span: Parallels in five cultures. Developmental Psychology 35: 466-477.

Muck, P. M., B. Hell und S. D. Gosling 2007: Construct Validation of a short Five-Factor Model Instrument – A self-peer study on the German adaptation of the Ten-Item Personality Inventory (TIPI-G). European Journal of Psychological Assessment 23: 166-175.

Ostendorf, F., 1990. Sprache und Persönlichkeitsstruktur. Zur Validität des Fünf-Faktoren-Modells der Persönlichkeit. Regensburg: Roderer.

Ostendorf, F. und A. Angleitner, 2004: NEO-Persönlichkeitsinventar nach Costa und McCrae, revidierte Fassung. Göttingen: Hogrefe.

Ozer, D. und V. Benet-Martínez, 2006: Personality and the prediction of consequential outcomes. Annual Review of Psychology 57: 401-421.

Rammstedt, B., 1997: Die deutsche Version des Big Five Inventory (BFI): Übersetzung und Validierung eines Fragebogens zur Erfassung des Fünf-Faktoren-Modells der Persönlichkeit. Unpublished thesis. University of Bielefeld, Germany.

Rammstedt, B., 2007a: The 10-Item Big Five Inventory (BFI-10): Norm values and investigation of socio-demographic effects based on a German population representative sample. European Journal of Psychological Assessment 23: 193-201.

Rammstedt, B., 2007b: Who worries and who is happy? Explaining individual differences in worries and satisfaction by personality. Personality and Individual Differences 43: 1626-1634.

Rammstedt, B. und O. P. John, 2005: Short version of the Big Five Inventory (BFI-K): Development and validation of an economic inventory for assessment of the five factors of personality. Diagnostica 51: 195-206.

Rammstedt, B. und O. P. John, 2007: Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of Research in Personality 41: 203-212.

Rammstedt, B., C. Kemper, M. C. Klein, C. Beierlein, und A. Kovaleva, 2012: Eine kurze Skala zur Messung der fünf Dimensionen der Persönlichkeit, Big-Five-Inventory-10 (BFI-10). Mannheim: GESIS.

Roberts, B. W., N. R. Kuncel, R. Shiner, A. Caspi und L. R. Goldberg, 2007: The power of personality: The comparative validity of personality traits, socioeconomic status and cognitive ability for predicting important life outcomes. Perspectives on Psychological Science 2:313-345.

Roberts, B. W., K. Walton, T. Bogg und A. Caspi, 2006: De-investment in work and non-normative personality trait change in young adulthood. European Journal of Personality 20: 461-474.

Romero, E., P. Villar, J. A. Gómez-Fraguela und L. López-Romero, 2012: Measuring personality traits with ultra-short scales: A study of the Ten Item Personality Inventory (TIPI) in a Spanish sample. Personality and Individual Differences 53: 289-293.

Schmidt, F. L. und J. E. Hunter, 1998: The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. Psychological Bulletin 124: 262-274.

Schupp, J. und J.-Y. Gerlitz, 2008: Das BFI-S: Big Five Inventory-SOEP. In A. Glöckner-Rist (Hg.): Zusammenstellung sozialwissenschaftlicher Items und Skalen. ZIS Version 12.00. Bonn: GESIS.

Selfhout, M., W. Burk, S. Branje, J. Denissen, M. van Aken und W. Meeus 2010: Emerging late adolescent friendship networks and Big Five personality traits: a social network approach. Journal of Personality 78: 509-538.

Sneed, C. D., 2002: Correlates and Implications for Agreeableness in Children. The Journal of Psychology 136: 59-67.

SOEP (Sozio-ökonomisches Panel) (2005). *Leben in Deutschland* [Datensatz und Codebook].
SOEP (Sozio-ökonomisches Panel) (2009). *Leben in Deutschland* [Datensatz und Codebook].
Srivastava, S., O. P. John, S. D. Gosling und J. Potter, 2002: Development of personality in early and middle adulthood: Set like plaster or persistent change? Journal of Personality and Social Psychology 84: 1041-1053.
Statistisches Bundesamt (2010). Statistik und Wissenschaft. Demographische Standards. Band 17. Wiesbaden: Statistisches Bundesamt.
Tupes, E. C. und R. E. Christal, 1961. Recurrent personality factors based on trait ratings. USAF Technical Report ASD-TR: 61-97.
Vassend, O. und A. Skrondal, 1995: Factor analytic studies of the NEO Personality Inventory and the five-factor model: The problem of high structural complexity and conceptual indeterminacy. Personality and Individual Differences 19: 135-147.
Vecchione, M., H. Schoen, J. L. González-Castro, J. Cieciuch, V. Pavlopoulos und G. V. Caprara, 2011: Personality correlates of party preference: The Big Five in five big European countries. Personality and Individual Differences 51: 737-742.
Viken, R. J., R. J. Rose, J. Kapiro und M. Koskenvuo, 1994: A developmental genetic analysis of adult personality: Extraversion and neuroticism from 18 to 59 years of age. Journal of Personality and Social Psychology 4: 722-730.
Walton, K. und B. W. Roberts, 2004: On the relationship between substance use and personality traits: Abstainers are not maladjusted. Journal of Research in Personality 38: 515-535.

| | |
|---|---|
| Anschrift der Autorin | Beatrice Rammstedt<br>GESIS – Leibniz-Institut für Sozialwissenschaften<br>Postfach 12 21 55<br>68072 Mannheim<br>E-Mail: beatrice.rammstedt@gesis.org |
| Ko-Autor/-innen | Constanze Beierlein<br>GESIS – Leibniz-Institut für Sozialwissenschaften<br>Mannheim<br><br>Christoph J. Kemper<br>Institut für medizinische und pharmazeutische<br>Prüfungsfragen, Mainz<br><br>Mira Céline Klein<br>Universität Mannheim<br><br>Anastassiya Kovaleva<br>Fakultät für Biologie, Universität Bielefeld |

# Appendix A

## *BFI-10*

Inwieweit treffen die folgenden Aussagen auf Sie zu?

| | | trifft überhaupt nicht zu | trifft eher nicht zu | weder noch | eher zutreffend | trifft voll und ganz zu |
|---|---|---|---|---|---|---|
| (1) | Ich bin eher zurückhaltend, reserviert. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| (2) | Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| (3) | Ich bin bequem, neige zur Faulheit. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| (4) | Ich bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| (5) | Ich habe nur wenig künstlerisches Interesse. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| (6) | Ich gehe aus mir heraus, bin gesellig. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| (7) | Ich neige dazu, andere zu kritisieren. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| (8) | Ich erledige Aufgaben gründlich. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| (9) | Ich werde leicht nervös und unsicher. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| (10) | Ich habe eine aktive Vorstellungskraft, bin fantasievoll. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |

methoden, daten, analysen · 2013, Jg. 7(2), S. 251-278 DOI: 10.12758/mda.2013.014

# Kurzskala zur Erfassung allgemeiner Selbstwirksamkeitserwartungen (ASKU)

# Short Scale for Measuring General Self-efficacy Beliefs (ASKU)

*Constanze Beierlein, Christoph J. Kemper, Anastassiya Kovaleva und Beatrice Rammstedt*

*Zusammenfassung*

Allgemeine Selbstwirksamkeitserwartungen beziehen sich auf die Einschätzung eigener Kompetenzen, Handlungen erfolgreich planen und ausführen zu können, um gewünschte Ziele zu erreichen. Selbstwirksamkeitserwartungen beeinflussen zahlreiche Aspekte alltäglicher Tätigkeiten. Empirische Studien verdeutlichen, dass sie mit einer Reihe sozialwissenschaftlicher Variablen aus den Bereichen Arbeit, Gesundheit und soziale Beziehungen in Verbindung stehen. Damit sind sie insbesondere für die interdisziplinäre Surveyforschung interessant, da sie einen Erklärungsbeitrag leisten können. Eine surveykompatible, d.h. ökonomische Kurzskala zur Erfassung des Konstrukts, fehlt jedoch bisher. Im Rahmen von drei empirischen Studien wurde die Allgemeine Selbstwirksamkeit Kurzskala („ASKU") mit drei Items entwickelt und validiert. Die Ergebnisse weisen darauf hin, dass die ASKU trotz ihrer Kürze das Konstrukt reliabel und valide misst. Darüber hinaus lassen die Befunde von Messinvarianzprüfungen darauf schließen, dass die Kurzskala in unterschiedlichen Erhebungsmodi ähnliche Messqualitäten aufweist. Die Skala zeigte

*Abstract*

General self-efficacy expectations refer to appraisals of one's own competencies to plan and execute actions in a successful way in order to achieve desired goals. Self-efficacy expectations have an impact on numerous aspects of day-to-day activities. Empirical studies show that they are linked to a number of key social scientific variables in the field of work, health, and social relations. This fact makes self-efficacy expectations particularly interesting for interdisciplinary survey research because they may explain variance in outcome variables. However, a short scale for assessing general self-efficacy expectations which is compatible for social surveys is still missing. In the scope of three empirical studies we developed and validated the "General Self-Efficacy Short Scale" (in German: ASKU) which contains only three items. The results indicate that despite of its brevity the ASKU measures the construct in a reliable and valid way. Furthermore, the findings suggest that the short scale show measurement invariance with respect to different assessment modes. As expected, the scale exhibits expected relations to sociodemographic

hypothesenkonforme Beziehungen zu sozio-demografischen Variablen sowie sozialwissenschaftlichen Inhaltsvariablen.

variables as well as social scientific content variables.

# 1 Einleitung[1]

Aus psychologischer Sicht wird das menschliche Handeln durch eine Vielzahl unterschiedlicher Variablen beeinflusst und gesteuert (z.B. Krampen 2000). Eine wichtige Rolle spielen hierbei Kognitionen (z.B. Wissen, Ziele) sowie Handlungsfähigkeiten. Bandura (1993) hat jedoch darauf aufmerksam gemacht, dass allein der Besitz von Wissen und Handlungsfähigkeiten noch nicht dazu führt, dass eine Person ein bestimmtes Verhalten auch tatsächlich zeigt. Es bedarf zusätzlich der Überzeugung, das Verhalten auch bei Eintreten von Hindernissen und Widerständen ausführen zu können.

Diesen Gedanken hat Bandura (1977) in seinem Konzept der „Selbstwirksamkeit" (engl. „Self-Efficacy") im Rahmen seiner sozialen Lerntheorie aufgegriffen. In diesem Konzept werden Menschen als Akteure verstanden, die aufgrund ihrer Fähigkeiten und Kompetenzen eine beachtliche Kontrolle ausüben und folglich darüber bestimmen, welche Erfahrungen sie im Leben machen und wie ihr Leben verläuft (Bandura 1977, 1982). Selbstwirksamkeitserwartungen beziehen sich dabei auf die Einschätzung der eigenen Fähigkeiten und Kompetenzen, Handlungen erfolgreich planen und ausführen zu können, um gewünschte Ziele zu erreichen (Bandura 1997; Pajares 1997). Das Konzept der Selbstwirksamkeit wurde auch in anderen theoretischen Ansätzen aufgegriffen: In seiner handlungstheoretischen Persönlichkeitspsychologie beschreibt Krampen (2000) Selbstwirksamkeitserwartungen als Situations-Handlungs-Erwartungen. Sie spiegeln damit die subjektive Einschätzung einer Person wider, ob der Person in einer konkreten Situation Handlungsmöglichkeiten zur Lösung von Problemen zur Verfügung stehen. Anders als Persönlichkeitseigenschaften beziehen sich Selbstwirksamkeitserwartungen damit nicht auf die Frage „wie bin ich?" sondern „wie gut kann ich etwas tun?" (Zimmerman/Cleary 2006: 47).

Die Selbstwirksamkeitserwartungen sind verbunden mit einem Gefühl von Kontrolle über die Umwelt sowie über das eigene Verhalten. Personen mit hohen Selbstwirksamkeitserwartungen haben den Eindruck, die Umwelt durch ihr Verhalten beeinflussen zu können. Selbstwirksamkeitserwartungen haben dadurch

1 Die Erstautorin bedankt sich bei Dr. Angelika Glöckner-Rist für methodische Hinweise.

einen bedeutenden Einfluss auf das Verhalten (Bandura 1997). In der Psychologie stellen sie einen zentralen motivationalen Prädiktor für Verhalten dar. Studienergebnisse verdeutlichen, auf welche Weise Selbstwirksamkeitserwartungen das Verhalten beeinflussen: So setzen sich Personen mit hohen Selbstwirksamkeitserwartungen z.B. anspruchsvollere Ziele, zeigen eine höhere Ausdauer und setzen effizientere Strategien beim Problemlösen ein (Bandura 1997; Pajares 1997; Schunk 1991). Selbstwirksamkeitserwartungen haben auch einen Effekt darauf, wie viel Stress und Anspannung Personen in risikoreichen oder stark herausfordernden Situationen erwarten. Personen mit höheren Selbstwirksamkeitserwartungen sind dann eher bereit, Risiken einzugehen, wenn sie mit ihrem Verhalten glauben, ein selbstgewähltes, erwünschtes Ziel zu erreichen (Breakwell 2007: 54-55). Dabei sind Selbstwirksamkeitserwartungen nicht statisch, sondern dynamisch: Sie können zum Beispiel durch Erfolgserfahrungen bzw. Lernprozesse beeinflusst werden (Bandura 1997).

Im *Gesundheitsbereich* werden hohe Selbstwirksamkeitserwartungen mit einer konstruktiven Stressbewältigung, einer günstigen Prognose bei der Erholung von Krankheiten sowie körperlichem und psychischem Wohlbefinden in Verbindung gebracht (z.B. Luszczynska/Benight/Cieslak 2009; Magaletta/Oliver 1999; Skaalvik/Skaalvik 2007). Studien zeigten zudem konsistente Ergebnisse dahingehend, dass hohe Selbstwirksamkeitserwartungen die Persistenz bezüglich Raucherentwöhnung und Alkoholabstinenz günstig beeinflussen (z.B. Gwalthney/Metrik/Kahler et al. 2009; Ilgen/McKellar/Tiet 2005; Sitharthan/Job/Kavanagh et al. 2003). Darüber hinaus gingen in Studien hohe Selbstwirksamkeitserwartungen mit erfolgreichen Copingstrategien im Umgang mit Schmerz einher (Pells/Shelby/Keffe et al. 2008; Turner/Ersek/Kemp 2005).

Im Hinblick auf *soziale Beziehungen* ließ sich zeigen, dass Selbstwirksamkeitserwartungen förderlich für prosoziales Verhalten sein können. Darüber hinaus stehen sie in einem positiven Zusammenhang mit der Lebenszufriedenheit (Caprara/Alessandri/Eisenberg 2012; Caprara/Steca 2005).

Im *pädagogischen Bereich* belegen Studien den Zusammenhang von Selbstwirksamkeitserwartungen und der Performanz in unterschiedlichen Kompetenzbereichen (z.B. Caprara/Fida/Vecchione et al. 2008; Williams/Williams 2010). So zeigte sich zum Beispiel in einer längsschnittlichen Studie, dass sich Selbstwirksamkeitserwartungen und Schulnoten wechselseitig positiv beeinflussen (Caprara/Vecchione/Alessandri et al. 2011). Selbstwirksamkeitserwartungen leisteten bei der Vorhersage akademischer Erfolge einen eigenständigen, zusätzlichen Erklärungsbeitrag, der über denjenigen von sozio-ökonomische Variablen oder der Intelligenz hinausgeht (Zuffianò/Alessandri/Gerbino et al. 2012).

Im Bereich *Arbeit und Organisationen* zeigen Befunde, dass Personen mit höheren Selbstwirksamkeitserwartungen eine höhere Resistenz gegenüber Arbeitsstress haben und ein stärkeres Commitment gegenüber ihrem derzeitigen Arbeitgeber ausdrücken (Akhtar/Saba/Adnan, in Druck; Heuven/Bakker/Schaufeli et al. 2006).

Auch im Bereich der *Politik* werden Selbstwirksamkeitserwartungen bereits zur Erklärung politischen Verhaltens herangezogen. Unter dem Fachbegriff „Political Efficacy" werden in Surveys individuelle Kompetenzerwartungen in politischen Handlungsfeldern erfasst (Vetter 1997). Diese bereichsspezifischen Selbstwirksamkeitserwartungen leisten einen substanziellen Beitrag zur Vorhersage politischer Beteiligungsbereitschaft (Vecchione/Caprara 2009).

Selbstwirksamkeitserwartungen stellen damit eine wertvolle Variable dar, die einen substanziellen Erklärungsbeitrag für sozialwissenschaftliche Inhaltsvariablen leisten kann. Den Empfehlungen Banduras (1997, 2006) folgend, werden Selbstwirksamkeitserwartungen häufig situations- oder kontextspezifisch erfasst; z.B. im Hinblick auf bestimmte Handlungsfelder wie das Lernen in der Schule oder auf konkrete Tätigkeiten wie das Autofahren. Bandura (2006) nimmt an, dass sich die Selbstwirksamkeitserwartungen von Menschen hinsichtlich verschiedener Handlungen unterscheiden. Demzufolge ist die Vorhersage eines spezifischen Verhaltens am besten möglich, wenn Selbstwirksamkeitserwartungen in Bezug auf dieses konkrete Verhalten erfasst werden. Für die Messung hätte dies jedoch zur Folge, dass jeweils handlungsspezifische Selbstwirksamkeitsskalen entwickelt und eingesetzt werden müssten. Aufgrund des zeitlichen und monetären Aufwands erscheint dies für viele sozialwissenschaftliche Surveys kaum realisierbar.

Theoretische Annahmen und empirische Befunde lassen jedoch darauf schließen, dass Selbstwirksamkeitserwartungen auch kontextübergreifend erfasst werden können (Caprara 2002): Die „allgemeine Selbstwirksamkeitserwartung" ist konzeptualisiert als eine über Situationen und Kontexte generalisierte Kompetenzerwartung (Bandura 2006: 307). Sie wird als stabile Erwartungshaltung betrachtet und bezieht sich dabei nicht auf ein spezifisches Handlungs- oder Funktionsfeld. Stattdessen spiegelt die allgemeine Selbstwirksamkeitserwartung wider, dass Menschen ihre Erfahrungen zu Erfolgen und Misserfolgen über Situationen hinweg verallgemeinern (Jerusalem/Schwarzer 1999).

Mehrere Studien zeigen, dass generalisierte Kompetenzerwartungen in unterschiedlichen Lebensbereichen positive Auswirkungen haben (z.B. Bandura 1997; Luszczynska/Gutiérrez-Dona/Schwarzer 2005). Die allgemeine Selbstwirksamkeit steht in positivem Zusammenhang mit Optimismus sowie der Arbeitszufriedenheit; negative Zusammenhänge zeigten sich unter anderem mit Ängstlich-

keit, Neigung zu Depressionen und Arbeitsstress (Luszczynska/Gutiérrez-Dona/Schwarzer 2005). Darüber hinaus stehen allgemeine Selbstwirksamkeitserwartungen in positiver Beziehung zur Arbeitsleistung (Judge/Bono 2001). Die allgemeine Selbstwirksamkeitserwartung wird deshalb auch als persönliche Bewältigungsressource aufgefasst (Schwarzer 1994).

Im Hinblick auf die Beziehung von Selbstwirksamkeitserwartungen zu den Big-Five-Persönlichkeitsfaktoren werden konsistente Befunde für drei der fünf Faktoren berichtet: In einer Metaanalyse von Judge und Ilies (2002) waren die Selbstwirksamkeitserwartungen moderat negativ mit Neurotizismus bzw. positiv mit Gewissenhaftigkeit assoziiert. Darüber hinaus fanden sich positive Beziehungen zur Extraversion. Der Zusammenhang der Selbstwirksamkeitserwartungen mit den restlichen beiden Big-Five-Dimensionen (Offenheit für Erfahrung/Verträglichkeit) ist nach Auskunft der beiden Autoren dagegen erst wenig beforscht. Judge und Ilies (2002) berichten jedoch für die Verträglichkeit die niedrigste Effektstärke.

Judge, Erez, Bono und Thoresen (2002) zufolge weist die allgemeine Selbstwirksamkeitserwartung Beziehungen zu verwandten Konstrukten wie dem Selbstwert (Rosenberg 1989) und den Kontrollüberzeugungen (Rotter 1966) auf. Alle drei dieser Konstrukte sowie der Neurotizismus werden als Aspekte eines übergeordneten Konstrukts, dem positiven Selbstkonzept, angesehen. Ein positives Selbstkonzept stellt eine sozial erwünschte Überzeugung dar. Der Zusammenhang zwischen sozialer Erwünschtheit und allgemeiner Selbstwirksamkeit konnte auch empirisch nachgewiesen werden: Eine Studie von Gravdal und Sandal (2006) ergab, dass Selbstwirksamkeitserwartungen positiv mit der Komponente der Selbsttäuschung des Konstrukts *Soziale Erwünschtheit* assoziiert sind.

## 2 Erfassung der allgemeinen Selbstwirksamkeit in sozialwissenschaftlichen Umfragen

Zur Erfassung der allgemeinen Selbstwirksamkeit wurden verschiedene Messinstrumente entwickelt (für einen Überblick siehe Scherbaum/Cohen-Charash/Kern 2006). In Deutschland hat sich insbesondere die von Schwarzer und Jerusalem vorgelegte Skala zur Erfassung der Allgemeinen Selbstwirksamkeitserwartung (SWE) etabliert. In der ursprünglichen Version enthielt die Skala 20 Items (Jerusalem/Schwarzer 1986). Heute wird in wissenschaftlichen Studien in der Regel eine

gekürzte Version der SWE von 10 Items eingesetzt (Jerusalem/Schwarzer 1999).[2] In der Skala werden allgemeine Selbstwirksamkeitserwartungen konzeptualisiert als „persönliche[n] Einschätzung der eigenen Kompetenzen, allgemein im täglichen Leben mit Schwierigkeiten und Barrieren zu Recht zu kommen und kritische Anforderungssituationen aus eigener Kraft erfolgreich bewältigen zu können" (Hinz/Schumacher/Albani et al. 2006: 26). Die Items beinhalten jeweils einen Hinweis auf ein Hindernis und die Einschätzung der Handlungsmöglichkeiten der Person, mit diesen Schwierigkeiten umzugehen. Die SWE als Messinstrument für die allgemeine Selbstwirksamkeitserwartung ist hinreichend reliabel und hat sich theoriekonform als eindimensional erwiesen (z.B. Leganger/Kraft/Roysamb 2000; Scholz/Gutiérrez-Doña/Sud et al. 2002). Die Skala wurde bereits für mehrere Sprachen adaptiert und validiert (Luszczynska/Gutiérrez-Dona/Schwarzer 2005; Luszczynska/Scholz/Schwarzer 2005).

Neben diesen positiven Eigenschaften weist die Skala jedoch insbesondere im Hinblick auf die Surveyforschung mehrere gravierende Nachteile auf. Zur Messung des eindimensionalen Konstrukts werden 10 Items benötigt, was laut Jerusalem und Schwarzer (1999) einer durchschnittlichen Bearbeitungszeit der Skala von 4 Minuten entspricht. Angesichts begrenzter zeitlicher und finanzieller Ressourcen in den meisten sozialwissenschaftlichen Surveys kann die Länge der Skala eine Hürde darstellen, sie in diesen Studien einzusetzen. Darüber hinaus ergeben sich bei der Skala Schwierigkeiten hinsichtlich der Formulierung einzelner Items: In einigen Items werden Begriffe verwendet, die einen großen Deutungsspielraum zulassen (z.B. „überraschende Ereignisse"; vgl. Faulbaum/Prüfer/Rexroth 2009). Zudem scheinen sich Inhalte einzelner Items zu wiederholen (z.B. Item 2: „Die Lösung schwieriger Probleme gelingt mir immer, wenn ich mich darum bemühe"; Item 8: „Für jedes Problem kann ich eine Lösung finden"). Des Weiteren fehlen Hinweise darauf, ob die Messäquivalenz der Skala über verschiedene Erhebungsmodi (z.B. „Papier-und-Bleistift", persönlich-mündliche Befragung) gewährleistet ist. Es wurde lediglich die Invarianz des Fragebogens über verschiedene Sprachversionen bzw. in verschiedenen Kulturen überprüft (Schwarzer/Bäßler/Kwiatek et al. 1997).

Aus den oben geschilderten Grenzen der von Jerusalem und Schwarzer (1999) entwickelten Skala, lassen sich die Forschungsziele des vorliegenden Beitrags ableiten: In Anlehnung an die von Jerusalem und Schwarzer (1999) konstruierte Skala soll eine kurze und reliable eindimensionale Skala zur Messung des Konstrukts *Allgemeine Selbstwirksamkeitserwartung* konstruiert und validiert werden.

2 Nähere Informationen zur 10-Item Skala von Jerusalem und Schwarzer (1999) finden sich auf folgender Website: http://userpage.fu-berlin.de/~health/germscal.htm (Zugriff am 15.12.2012).

Die Kurzskala soll dabei zur Erfassung der allgemeinen Selbstwirksamkeitserwartung in der deutschsprachigen Allgemeinbevölkerung ab 18 Jahren dienen und in unterschiedlichen Erhebungsmodi eingesetzt werden können. Die Messäquivalenz der Skala für verschiedene Erhebungsmodi soll sichergestellt werden. Die Kurzskala soll einen Erklärungsbeitrag für die Vorhersage sozialwissenschaftlicher Inhaltsvariablen aus den Bereichen Gesundheit, Arbeit und soziale Beziehungen leisten.

## 3 Methode

### *Stichproben*

Im Rahmen der Konstruktion und Validierung der Skala im GESIS-Projekt „Entwicklung einer Standardbatterie für psychologische Merkmale in sozialwissenschaftlichen Umfragen" wurden verschiedene Stichproben erhoben. Die Charakteristika dieser Stichproben sind in Tabelle 1 dargestellt. *Stichprobe 1* ist eine Quotenstichprobe ($N$ = 539), die nach den Merkmalen Geschlecht, Alter, Bildung und Bundesland geschichtet ist. Die Grundgesamtheit war definiert als „alle in der Bundesrepublik Deutschland in Privathaushalten lebenden deutschsprachigen Personen ab 18 Jahren". Die Erhebung, welche neben den Selbstwirksamkeitsitems weitere Skalen beinhaltete, erfolgte in zwei Wellen mit einem zeitlichen Abstand von 6 bis 10 Wochen. An Welle 2 nahmen $N$ = 338 Befragungspersonen der Welle 1 teil. Die Daten wurden im Rahmen eines Interviews (CAPI; Computer Assisted Personal Interview) oder durch die Vorgabe eines Papierfragebogens („Papier-und-Bleistift") erhoben. Der Anteil der Befragungen im jeweiligen Erhebungsmodus wurde vor der Feldphase in einem Plan festgelegt. Dieser sah bestimmte Kontingente für beide Erhebungsmodi vor. Danach wurde bei der Mehrzahl der Befragten die Befragung im CAPI-Modus durchgeführt. Die Zuordnung der Personen zu den Bedingungen in Welle 1 und 2 erfolgte dann auf Zufallsbasis. Die Erhebung aller in der Umfrage erfassten Konstrukte dauerte insgesamt im Mittel 53 Minuten ($SD$ = 12). Bei *Stichprobe 2* handelt es sich ebenfalls um eine Quotenstichprobe, geschichtet nach Geschlecht, Alter und Bildung ($N$ = 359). Die Datenerhebung erfolgte jedoch anders als in Stichprobe 1 über das Internet (CAWI). Grundgesamtheit waren die Teilnehmer eines Online-Access-Pools im Alter von 18 Jahren und älter, die in Deutschland leben. Die Bearbeitung des gesamten Onlinefragebogens dauerte im Durchschnitt 23 Minuten ($SD$ = 8). *Stichprobe 3* mit $N$ = 1.134 Befragungspersonen ist eine Zufallsstichprobe, die repräsentativ für die Wohnbevölkerung in Deutschland über einem Alter von 18 Jahren ist. Sie wurde mithilfe des ADM-Stichprobensystems F2F („Random Route") der Arbeitsgemeinschaft deutscher Marktforschungsinstitute

Tabelle 1 *Charakteristika der Stichproben*

| | | Stichprobe 1 | | Stichprobe 2 | Stichprobe 3 |
|---|---|---|---|---|---|
| | | Welle 1 | Welle 2 | | |
| *Stichprobe* | | | | | |
| Umfang [*N*] | | 539 | 338 | 359 | 1.134 |
| Art | | Quote | Quote | Quote | Zufall |
| Modus | | CAPI, P&B | CAPI, P&B | CAWI | CAPI, CASI |
| *Zusammensetzung* | | | | | |
| Geschlecht [% Frauen] | | 52,5% | 52,1% | 51,2% | 55,6% |
| Alter [*M*(*SD*)] | | 47.2 (15.2) | 46.7 (15.1) | 48.3 (13.0) | 53.3 (18.4) |
| Bildung | ≤ 9 Jahre | 44,7% | 45,3% | 40,4% | 37,2% |
| | 10 Jahre | 30,2% | 27,9% | 31,8% | 37,0% |
| | ≥ 11 Jahre | 23,7% | 25,4% | 27,9% | 25,8% |

*Anmerkungen: CAPI = Computer Assisted Personal Interview, CASI = Computer Assisted Self Interview, CAWI = Computer Assisted Web Interview, P&B = Papier-und-Bleistift (Selbstausfüller).*

gezogen. Die Daten dieser Interviews wurden größtenteils im CAPI-Modus erhoben; der letzte Teil der Erhebung im CASI-Modus (CASI: Computer Assisted Self Interview). Die Dauer der Befragung lag durchschnittlich bei 43 Minuten (*SD* = 13). Tabelle 1 fasst die Charakteristika der drei Stichproben zusammen.

### *Vorgehen*

Die Itemselektion und die Validierung erfolgten in einem mehrstufigen Verfahren auf der Basis qualitativer und quantitativer Analysen. Den Ausgangspunkt für die Itemselektion stellte die von Jerusalem und Schwarzer (1999) vorgelegte 10-Item-Skala (SWE) dar. In einem ersten Schritt wurde eine Reanalyse von Daten aus 5 heterogenen Stichproben (Erwachsene, Studierende, Personen mit und ohne Migrationshintergrund; *N* = 2.115; Alter: *M* = 33.4, *SD* = 17.1; 54% weiblich) durchgeführt; die Daten hierzu wurden von Jerusalem und Schwarzer auf folgender Internetseite zur Verfügung gestellt: Quelle: http://www.selbstwirksam.de/. Ziel war es, die psychometrischen Eigenschaften der SWE-Items zu erkunden und eine erste Vorauswahl zu treffen. Als Kriterien wurden die statistischen Eigenschaften der Items inklusive ihrer faktoriellen Struktur herangezogen. Ausgewählt wurden diejenigen Items, die in einer Hauptachsenfaktorenanalyse substanziell und auf einem gemeinsamen Faktor luden und keine bis geringe Nebenladungen aufwiesen.

Zusätzlich wurde die Höhe der Itemtrennschärfe der Items berücksichtigt, wobei diejenigen der Items mit einer höheren Itemtrennschärfe bevorzugt wurden.

Die im Rahmen der Reanalyse ausgewählten sieben Items wurden im Anschluss in einem Expertenreview leicht modifiziert (Prüfer/Rexroth 2000). Die Modifikation der Items hatte zum Ziel, die Verständlichkeit der Items für alle Teile der Bevölkerung zu erhöhen. Dabei wurde insbesondere die Satzstruktur vereinfacht. Darüber hinaus ergaben die Reanalysen, dass die von Jerusalem und Schwarzer (1999) vorgeschlagene und in den analysierten Studien verwendete vierstufige Antwortskala die Differenzierbarkeit der Antworten im oberen Bereich der Skala reduzierte (vgl. Bandura 1997; Pajares/Hartley/Valiante 2001). Dieser Befund entspricht den Ergebnissen aus früheren Untersuchungen: In mehreren Studien wurde eine negative Schiefe der Skalenwerteverteilung und damit eine geringere Verteilungsbreite der Werte am oberen Ende der Skala berichtet (Scholz/Gutiérrez-Dona/Sud et al. 2002; Schwarzer/Born 1997). Um diesem Problem eines Deckeneffekts zu begegnen, wurde die von Jerusalem und Schwarzer verwendete vierstufige Antwortskala durch eine fünfstufige ersetzt.

Die sieben leicht modifizierten Items wurden dann einem kognitiven Pretest unterzogen ($N$ = 20; Alter $M$ = 46.2, $SD$ = 14.7; 50% weiblich). Kognitive Pretests haben zum Ziel, das Verständnis von Items, Fragen und Antwortvorgaben in einem Fragebogen zu überprüfen (siehe auch Prüfer/Rexroth 2000). Eine zentrale Erkenntnis aus dem Pretest war, dass manche der in den Items verwendeten Begriffe zu abstrakt waren und unterschiedliche Interpretationen erlaubten (z.B. „unerwartete Situationen", „Widerstände"). Folglich wurde eine weitere sprachliche Modifikation der Items vorgenommen. Der Pretest ergab darüber hinaus, dass Items der SWE-Skala als sehr ähnlich empfunden werden. Wie bereits vor dem Pretest erwartet, schilderten die befragten Personen den Eindruck, dass sich einige Items wiederholten. Dieses Ergebnis sprach für eine weitere Reduktion der Itemanzahl. Ein Item wurde ausgeschlossen, da die Formulierung wenig konkret und zu allgemein optimistische Lebenserwartungen erfasste („Was auch immer passiert, ich werde schon klarkommen.").

Im Anschluss wurden sechs Items im Rahmen von *Stichprobe 1* einem empirischen Test unterzogen. Eine weitere Reduktion der Itemanzahl wurde vorgenommen, in dem diejenigen Items ausgewählt wurden, die in einer Hauptachsenfaktorenanalyse am höchsten auf dem gemeinsamen Faktor luden und hohe Itemtrennschärfen aufwiesen. Darüber hinaus wurde darauf geachtet, sprachlich-inhaltliche Wiederholungen bei den Items zu vermeiden. Nach dieser weiteren Reduktion der Itemanzahl wurden drei Items für die finale Version der „Allgemeinen Selbstwirksamkeitskurzskala" (ASKU) ausgewählt.

Es wurde eine 5-stufige Antwortskala mit den folgenden Kategorienbezeichnungen gewählt: 1) „trifft gar nicht zu", 2) „trifft wenig zu", 3) „trifft etwas zu", 4) „trifft ziemlich zu", 5) „trifft voll und ganz zu" (vgl. Rohrmann 1978). Der individuelle Skalenmittelwert der Items ergibt sich dabei aus der Summe der Antworten auf den drei Items geteilt durch die Itemanzahl. Die Items und Antwortskala der ASKU sind im Appendix aufgeführt.

Die psychometrischen Eigenschaften der ASKU wurden in Stichprobe 2 und 3 überprüft. Hierzu wurden neben der Originalskala von Jerusalem und Schwarzer (1999) sozialwissenschaftliche Inhaltsvariablen aus den Bereichen Arbeit, Gesundheit, Politik und soziale Beziehungen herangezogen. Darüber hinaus wurden psychologische Konstrukte, mit denen die Selbstwirksamkeit auf der Grundlage theoretischer Annahmen und empirischer Befunde in Verbindung steht, ebenfalls in die Untersuchung aufgenommen.

### *Statistische Datenanalyse*

Die quantitativen Analysen zur Itemselektion umfassten neben der Betrachtung deskriptiver Indikatoren der Item- und Skalenstatistiken (Mittelwert, Standardabweichung, Spannweite, Trennschärfe, Schwierigkeit, Schiefe, Exzess) exploratorische Faktorenanalysen (PAF; Hauptachsenanalysen) mit der Statistiksoftware SPSS 19.0. Die faktorielle Struktur der neuen Kurzskala wurde mittels konfirmatorischer Faktorenanalysen (CFA; vgl. Brown 2006; Kline 2011) überprüft. Als Gütekriterien für den Modellfit wurden dabei folgende Indizes herangezogen (vgl. Beauducel/Wittmann 2005; Brown 2006; Browne/Cudeck 1993; Hu/Bentler 1999): $\chi^2$ (*df*, *p*), Comparative Fit Index (CFI; zufriedenstellender Fit > .95), Tucker-Lewis-Index (TLI; zufriedenstellender Fit > .95), Root Mean Square Error of Approximation (RMSEA; zufriedenstellender Fit < .08).

Die Messinvarianzprüfungen hinsichtlich verschiedener Erhebungsmodi wurden auf der Basis einer „Multiple Group Confirmatory Factor Analysis" (MGCFA) durchgeführt (Brown 2006). Aufgrund des Erhebungsdesigns der Studien konnten drei Gruppen, in denen die Skala in unterschiedlichem Erhebungsmodus (CAPI, Papier-und-Bleistift, Online) getestet wurde, miteinander verglichen werden. Zur Parameterschätzung wurde das Maximum-Likelihood-Verfahren eingesetzt. In Anlehnung an Byrne, Shavelson und Muthén (1989; siehe auch Byrne 2009; Vandenberg/Lance 2000) wurde im Rahmen der Messinvarianzprüfung ein schrittweises Vorgehen gewählt. Dabei wurden drei Modelle mit unterschiedlichen Parameterrestriktionen getestet: 1) *Konfigurale Invarianz*: über die Gruppen hinweg wurde ein Modell spezifiziert, welches das gleiche Muster an fixierten und freien Faktorladungen beeinhaltete. 2) *Metrische Invarianz*: Die Faktorladungen wurden

Tabelle 2 Itemstatistiken der *vorläufigen* Items der Allgemeinen Selbstwirksamkeit Kurzskala (ASKU) in Stichprobe 1 getrennt nach Messzeitpunkt (MZP) und Erhebungsmodus (Modus)

| | MZP | Modus | *M* | *SD* | $r_{it}$ | Schiefe | Kurtosis |
|---|---|---|---|---|---|---|---|
| 1. In vielen schwierigen Situationen kann ich mich auf meine Fähigkeiten verlassen. | 1 | CAPI | 3.98 | 0.78 | .65 | -0.63 | 0.61 |
| | 1 | P&B | 3.75 | 0.73 | .59 | -0.18 | -0.17 |
| | 2 | CAPI | 4.02 | 0.80 | .67 | -0.68 | 0.54 |
| | 2 | P&B | 3.86 | 0.74 | .63 | -0.18 | -0.28 |
| 2. Die meisten Probleme kann ich aus eigener Kraft gut meistern. | 1 | CAPI | 3.86 | 0.82 | .68 | -0.43 | 0.04 |
| | 1 | P&B | 3.68 | 0.73 | .56 | -0.28 | -0.04 |
| | 2 | CAPI | 3.89 | 0.78 | .69 | -0.49 | 0.08 |
| | 2 | P&B | 3.69 | 0.82 | .70 | -0.80 | 1.20 |
| 3. Auch schwierige Aufgaben kann ich in der Regel gut lösen. | 1 | CAPI | 3.83 | 0.81 | .71 | -0.31 | -0.22 |
| | 1 | P&B | 3.66 | 0.78 | .59 | 0.01 | -0.46 |
| | 2 | CAPI | 3.82 | 0.81 | .72 | -0.62 | 0.89 |
| | 2 | P&B | 3.66 | 0.79 | .77 | -0.31 | 0.36 |

*Anmerkungen: M = Mittelwert, SD = Standardabweichung, $r_{it}$ = Trennschärfe. CAPI = Computer-Assisted Personal Interview; P&B = Papier & Bleistift. Stichprobenumfang Messzeitpunkt 1: CAPI N ≥ 404; P&B N ≥ 130; Stichprobenumfang Messzeitpunkt 2: CAPI N ≥ 224; P&B N = 111. Fünfstufiger Antwortmodus: 1 = trifft gar nicht zu, 2 = trifft wenig zu, 3 = trifft etwas zu, 4 = trifft ziemlich zu, 5 = trifft voll und ganz zu. Der Range der Itemscores betrug zu allen Messzeitpunkten bzw. allen Erhebungsmodi 1 bis 5.*

für alle Gruppen gleichgesetzt. 3) *Skalare Invarianz*: Die Intercepts der manifesten Variablen wurden über die Gruppen hinweg gleichgesetzt. Da es sich bei den drei Modellen um eingebettete („nested") Modelle handelt, wurden die Modelle mittels $\chi^2$-Differenz-Tests miteinander verglichen. Darüber hinaus gelten nach den Richtlinien von Cheung und Rensvold (2002) Veränderungen der Modellgüte gemessen über den CFI um .01 oder weniger als Hinweis darauf, die Invarianzhypothese nicht zurückzuweisen.

Die konfirmatorischen Faktorenanalysen sowie die MGCFA wurden mit dem Programm M*plus* 6.1 (Muthén/Muthén 1998-2010) durchgeführt.

# 4 Ergebnisse

### *Item- und Skalenstatistiken*

Tabelle 2 und 3 geben die Ergebnisse der Item- und Skalenstatistiken für die vorläufige sowie für die endgültige Version der Items der ASKU wider. Die Ergeb-

Tabelle 3 Itemstatistiken der *finalen* Items der Allgemeinen Selbstwirksamkeit Kurzskala (ASKU) in Stichprobe 2 (CAWI, $N$ = 359) und Stichprobe 3 (CAPI, $N$ = 1.134)

| | Modus | $M$ | $SD$ | $r_{it}$ | Schiefe | Kurtosis |
|---|---|---|---|---|---|---|
| 1. In schwierigen Situationen kann ich mich auf meine Fähigkeiten verlassen. | CAWI | 3.83 | 0.87 | .74 | -0.71 | 0.71 |
| | CAPI | 4.06 | 0.86 | .67 | -0.79 | 0.41 |
| 2. Die meisten Probleme kann ich aus eigener Kraft gut meistern. | CAWI | 3.87 | 0.82 | .74 | -0.65 | 0.62 |
| | CAPI | 4.04 | 0.81 | .71 | -0.73 | 0.58 |
| 3. Auch anstrengende und komplizierte Aufgaben kann ich in der Regel gut lösen. | CAWI | 3.79 | 0.86 | .73 | -0.67 | 0.58 |
| | CAPI | 3.88 | 0.90 | .70 | -0.72 | 0.42 |

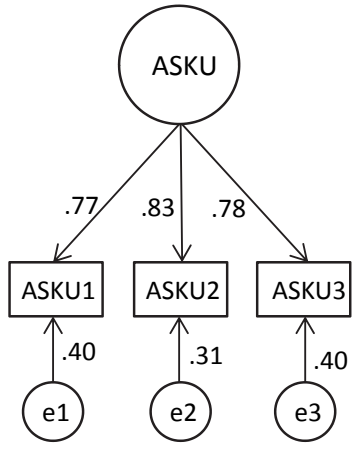
*Anmerkungen: CAPI = Computer Assisted Personal Interview, CAWI = Computer Assisted Web Interview. M = Mittelwert, SD = Standardabweichung, $r_{it}$ = Trennschärfe. Fünfstufiger Antwortmodus: 1 = trifft gar nicht zu, 2 = trifft wenig zu, 3 = trifft etwas zu, 4 = trifft ziemlich zu, 5 = trifft voll und ganz zu. Der Range der Itemscores betrug zu allen Messzeitpunkten bzw. allen Erhebungsmodi 1 bis 5.*

nisse zur vorläufigen Version der Items basieren auf der Erhebung in Stichprobe 1 (Papier-und-Bleistift, CAPI); die Ergebnisse zur finalen Version der Items auf den Erhebungen in Stichproben 2 (CAWI) und 3 (CAPI). Die Items weisen durchgehend zufriedenstellende Trennschärfen von über $r_{it}$ = .50 auf. Mit Mittelwerten zwischen $M$ = 3.66 und 4.06 sowie Standardabweichungen zwischen $SD$ = 0.73 und 0.90 konzentrieren sich die Werte im oberen Bereich der Antwortskala. Jedoch sind keine Verteilungsauffälligkeiten in Bezug auf Schiefe und Kurtosis festzustellen.

### *Reliabilität und Stabilität*

Die Messgenauigkeit der ASKU wurde im Rahmen von Strukturgleichungsmodellen (SEM; Jöreskog 1969) auf Grundlage der Ladungen und Fehlervarianzen aus den Messmodellen in den drei Stichproben geschätzt. Als Schätzer wurde der Koeffizient ω von McDonald (1999: 90) verwendet. Der Koeffizient gibt das Ausmaß an, in dem eine latente Variable (d.h. ein Konstrukt) von den Items geteilte Varianz reflektiert (Krohne/Hock 2007). Laut Schweizer (2011) ist dieser Schätzer der Reliabilität besser geeignet als Cronbach α. Die Interpretation der Höhe von McDonald ω ist analog zu Cronbach α. Die Schätzer der Reliabilität für die ASKU wurden anhand der gleichgesetzten Ladungen der drei Items auf dem gemeinsamen Faktor ermittelt und betragen in Stichprobe 1 ω = .81 für Welle 1 und ω = .84 für Welle 2, in Stichprobe 2 ω = .86 und in Stichprobe 3 ω = .84. Demnach liegt die Reliabilität der ASKU zwischen .81 und .86. Dies entspricht einer für Gruppenuntersuchungen ausreichenden Reliabilität. Gegenüber der 10-Item-Skala von Jerusalem und

Abbildung 1 Faktorenmodell der allgemeinen Selbstwirksamkeit (ASKU) mit standardisierten Ladungen in Stichprobe 3




Schwarzer (1999; α = .92; ω = .92) ging eine Kürzung auf drei Items somit nur mit einer marginalen Reduktion der Reliabilität einher. Neben McDonald ω wurde auch die Stabilität der ASKU-Skalenwerte durch eine Korrelation in den beiden Wellen von Stichprobe 1 ermittelt. Die Stabilität liegt bei $r_{tt}$ = .50. Die Stabilität wurde dabei auf der Basis der vorläufigen Itemformulierungen berechnet.

### *Faktorielle Validität*

Die faktorielle Validität der ASKU wurde separat für alle drei Stichproben mittels konfirmatorischer Faktorenanalysen überprüft. Getestet wurde ein Modell, bei dem die Faktorladungen der drei Items auf dem gemeinsamen Faktor gleichgesetzt wurden, um den Modellfit vergleichen zu können. In den drei Stichproben bzw. den drei Erhebungsmodi erreichten die Items der ASKU jeweils standardisierte Faktorladungen von .77 und höher. Die globale Modellgüte kann als zufriedenstellend bewertet werden (Stichprobe 1 [Welle1, Papier-und-Bleistift]: $\chi^2$ = 0.79, *df* = 2, *p* = 0.67; CFI = 1.00; TLI = 1.00; RMSEA = .001; Stichprobe 2 [CAWI]: $\chi^2$ = 0.48, *df* = 2, *p* = 0.78; CFI = 1.00; TLI = 1.00; RMSEA = .01; Stichprobe 3 [CAPI]: $\chi^2$ = 9.54, *df* = 2, *p* = 0.01; CFI = .99; TLI = .99; RMSEA = .06). Diese Ergebnisse lassen auf die faktorielle Validität der Kurzskala schließen. Abbildung 1 stellt exemplarisch das Faktorenmodell der ASKU mit den standardisierten Ergebnissen aus Stichprobe 3 dar.

Tabelle 4 Modellfitindizes der Invarianzmodelle für Gruppen mit unterschiedlichen Erhebungsmodi der ASKU (Stichprobe 1: CAPI, $n_1$ = 407, P&B, $n_2$ = 131, Stichprobe 2: CAWI $n_3$ = 359) sowie Ergebnisse des $\chi^2$-Differenztests

| | $\chi^2$ | *df* | *p* | CFI | TLI | RMSEA | Konfidenz-interval | $\chi^2_{diff}$ | $df_{diff}$ | *p* |
|---|---|---|---|---|---|---|---|---|---|---|
| *CAPI, P&B* (Stichprobe 1, Welle 1) | | | | | | | | | | |
| 1. Konfigurale Invarianz | 0 | 0 | 0 | 1.00 | 1.00 | 0 | (0;0) | - | - | - |
| 2. Metrische Invarianz | 0.87 | 2 | 0.65 | 1.00 | 1.00 | .001 | (0;.09) | - | - | - |
| 3. Skalare Invarianz | 2.57 | 4 | 0.63 | 1.00 | 1.00 | .001 | (0;.08) | 1.70 | 2 | .43 |
| *CAPI, P&B* (Stichprobe 1), CAWI (Stichprobe 2) | | | | | | | | | | |
| 1. Konfigurale Invarianz | 0 | 0 | 0 | 1.00 | 1.00 | 0 | (0;0) | - | - | - |
| 2. Metrische Invarianz | 4.02 | 4 | 0.40 | 1.00 | 1.00 | .004 | (0;.09) | - | - | - |
| 3. Skalare Invarianz | 15.64 | 8 | 0.05 | 0.99 | 0.99 | .06 | (.01;.09) | 11.62 | 4 | .02 |

*Anmerkungen. CAPI = Computer Assisted Personal Interview, CASI = Computer Assisted Self Interview, CAWI = Computer Assisted Web Interview, P&B = Papier-und-Bleistift (Selbstausfüller).*

### *Messinvarianzprüfungen*

Es wurden zwei Messinvarianzprüfungen durchgeführt: (1) Zunächst wurden die Erhebungsmodi CAPI und Papier-und-Bleistift in Stichprobe 1 verglichen. Hierfür wurden lediglich die Daten aus Welle 1 herangezogen. Damit sollte der Einfluss von Messwiederholungseffekten ausgeschlossen werden. (2) Die zweite Messinvarianzprüfung richtete sich auf den Vergleich der drei über die Studien hinweg realisierten Erhebungsmodi CAPI, P&B und CAWI. Hierfür wurden Daten aus Stichprobe 1 und 2 herangezogen. Dabei ist zu beachten, dass sich die Itemformulierungen zwischen den beiden Stichproben geringfügig unterschieden (siehe Tabelle 2 und 3). Stichprobe 3 wurde aufgrund des stark abweichenden Stichprobenumfangs nicht berücksichtigt. Um der latenten Variablen eine Skala zuzuweisen, wurde in dem zu testenden Modell die Ladung des ersten Items (ASKU1) auf 1 fixiert. Die Ergebnisse der beiden Messinvarianzprüfungen sind in Tabelle 4 dargestellt.

Bei der Interpretation der Ergebnisse muss zunächst folgendes Spezifikum der Modelltests berücksichtigt werden: Das zu prüfende einfaktorielle Modell war mit drei Indikatoren gerade identifiziert. Auf eine Gleichsetzung der Itemladungen wurde im Rahmen der Messinvarianzprüfung verzichtet. Von einer Interpretation des Modellfits im Rahmen des Tests auf konfigurale Invarianz wurde deshalb abgesehen. Auf die Interpretation des Ergebnisses des $\chi^2$-Differenztests zwischen dem konfiguralen und dem metrischen Modell wurde ebenfalls verzichtet.

Der Vergleich zwischen den beiden Erhebungsmodi CAPI und P&B in Stichprobe 1 zeigt, dass sich der Modellfit mit zunehmenden Parameterrestriktionen nicht substanziell bzw. statistisch signifikant verringert. Die Ergebnisse deuten darauf hin, dass die Hypothese der äquivalenten Faktorladungen und Itemintercepts nicht zurückgewiesen werden kann. Der $\chi^2$-Differenztest wird beim Vergleich der Modelle der metrischen und der skalaren Invarianz nicht statistisch signifikant. In weiteren Analysen werden die Daten der beiden Teilstichproben (CAPI, P&B) deshalb zusammengefasst.

Wie die Ergebnisse für den Vergleich der drei Erhebungsmodi verdeutlichen, weist das Modell zur Invarianz einen zufriedenstellenden Modellfit auf. Der $\chi^2$-Differenztest wird nicht statistisch signifikant. Im Vergleich zum restriktiveren Modell (skalare Invarianz) muss jedoch konstatiert werden, dass das restriktivere Modell eine schlechtere Passung an die empirischen Daten zeigt. Dass die Modelle statistisch signifikant voneinander abweichen, belegen auch die Ergebnisse des $\chi^2$-Differenztest. Das bedeutet, dass sich die Messungen in den Gruppen nicht hinsichtlich der Faktorladungen unterscheiden. Über die drei Erhebungsmodi hinweg variieren jedoch die Itemschwierigkeiten, so dass die Hypothese einer skalaren Messinvarianz des Instruments verworfen werden muss. Es zeigte sich, dass die Itemschwierigkeiten in der Onlinebefragung sowie in der Papier-und-Bleistiftbefragung unterhalb der Itemschwierigkeiten in der CAPI-Befragung lagen. Dies bedeutet, dass die Items in der CAPI-Befragung leichter im Sinne des Konstrukts beantwortet wurden als in den anderen beiden Bedingungen.

### *Konstruktvalidität*

Im Rahmen der Validierung der ASKU wurden die theoretisch erwarteten Beziehungen der neuen Kurzskala mit (weiteren) sozialwissenschaftlichen Variablen überprüft. Zunächst wurde die Korrelation der ASKU mit der 10-Item-Skala von Schwarzer und Jerusalem (1999) bestimmt, einem alternativen Maß für das Konstrukt allgemeine Selbstwirksamkeitserwartung. Der Zusammenhang der beiden Skalen betrug in Stichprobe 1 $r = .75$ ($p = .001$); damit zeigt die ASKU eine hohe konvergente Validität mit dieser bereits etablierten Skala. Anschließend wurde die ASKU mit weiteren aus der Fachliteratur bekannten typischen Korrelaten der allgemeinen Selbstwirksamkeit in Beziehung gesetzt. Tabelle 5 stellt die Ergebnisse der Korrelationsanalysen dar.

Wie aufgrund der Befunde von Judge und Bono (2001) sowie Judge et al. (2002) zu erwarten, korrelierte die neu konstruierte ASKU im Sinne der Effektstärkenbeurteilung nach Cohen (1992) stark positiv mit der internalen Kontrollüberzeugung und dem Selbstwert. In Übereinstimmung mit den theoretischen Annah-

men sowie empirischen Befunden zeigte die Skala starke positive Beziehungen zu Optimismus, der allgemeinen Lebenszufriedenheit, sowie schwache bis mittelstarke Korrelationen mit der Partnerschafts- und der Arbeitszufriedenheit (vgl. z.B. Bandura 1997; Hinz/Schumacher/Albani et al. 2006; Judge/Bono 2001; Luszczynska/Gutiérrez-Dona/Schwarzer 2005). Hohe Selbstwirksamkeitserwartungen gingen in Stichprobe 3 zudem mit einer erhöhten Risikobereitschaft einher (vgl. Breakwell 2007): Personen, die von ihrer eigenen Problemlösekompetenz auch in neuartigen Situationen überzeugt waren, sprechen sich diese Fähigkeiten auch in riskanten Situationen zu. Selbstwirksamkeitserwartungen standen in Stichprobe 3 auch mit der kristallinen Intelligenz in positiver Beziehung: Das im Laufe des Lebens erworbene Wissen sowie die angemessene Anwendung von Problemlösestrategien gehen demzufolge mit Kompetenzerwartungen einher. Dies entspricht empirischen Befunden, wonach sich Kompetenzerleben und Selbstwirksamkeitserwartungen wechselseitig positiv beeinflussen (Bandura 1997). Selbstwirksamkeitserwartungen waren in der vorliegenden Studie zudem mit einer stärkeren Neigung zur Übertreibung positiver Qualitäten im Sinne eines sozial erwünschten Antwortverhaltens assoziiert (vgl. Gravdal/Sandal 2006).

Im Hinblick auf die fünf Hauptfaktoren der Persönlichkeit (Fünf-Faktoren-Modell) war die ASKU in den drei Stichproben in den jeweils erwarteten Richtungen mittelstark mit Neurotizimus, Extraversion, Gewissenhaftigkeit und Offenheit assoziiert (vgl. Luszczynska/Gutiérrez-Doña/Schwarzer 2005; Zuffianò/Alessandri/Gerberino et al. 2012). Die Korrelationsstärke mit Offenheit fiel dabei höher aus als aufgrund der Metaanalyse von Judge und Ilies (2002) erwartet. Erwartungsgemäß zeigte sich kein Zusammenhang der Big-Five-Persönlichkeitsdimension Verträglichkeit mit der allgemeinen Selbstwirksamkeit.

In Tabelle 5 sind die Korrelationen der bereits etablierten 10-Item Skala von Schwarzer und Jerusalem (1999) denjenigen der neuen ASKU gegenübergestellt. Hierzu wurden Signifikanztests auf der Basis der nach Fisher Z transformierten Korrelationskoeffizienten der ASKU und der SWE durchgeführt. Beide Skalen wurden im Rahmen von Stichprobe 1 erhoben. Statistisch signifikante Abweichungen der beiden Korrelationskoeffizienten zeigten sich für die Variablen Optimismus, Pessimismus, Lebenszufriedenheit sowie die Big Five Dimension Offenheit für Erfahrungen. Die Richtung der Korrelationen beider Skalen stimmte jedoch bei allen aufgeführten psychologischen Konstrukten stets überein.

Tabelle 5 Validitätskoeffizienten der bereits etablierten Skala SWE (Schwarzer & Jerusalem, 1999) und der ASKU (Beierlein u.a. 2012) mit psychologischen Variablen sowie Ergebnis der Fisher Z-Tests auf Gleichheit der Korrelationskoeffizienten von SWE und ASKU (S1 = Stichprobe 1; S2 = Stichprobe 2, S3 = Stichprobe 3)

| | | S1 | | | S2 | S3 |
|---|---|---|---|---|---|---|
| Psychologische Variable | Messverfahren | SWE | ASKU | Fisher *Z* | ASKU | ASKU |
| Kontrollüberzeugung - Internal | S1: Jakoby/Jacob (1999); S3: Kovaleva u.a. (2012) | .71** | .73** | -1.01 | - | .62** |
| Kontrollüberzeugung - External | S1: Jakoby/Jacob (1999); S3: Kovaleva u.a. (2012) | -.41** | -.37** | -1.44 | - | -.34** |
| Optimismus | S1: Glaesmer u.a. (2008); S3: Kemper u.a. (in Druck) | .54** | .60** | -2.47** | - | .49** |
| Pessimismus | S1: Glaesmer u.a. (2008); S3: Kemper u.a. (in Druck) | -.47** | -.30** | -6.11** | - | -.40** |
| Lebenszufriedenheit | S1: Diener u.a. (1985) S3: in Anlehnung an SOEP | .61** | .49** | 4.85** | - | .41** |
| Big Five – Neurotizismus | Rammstedt/John (2007) | -.44** | -.44** | 0 | -.37 | -.31** |
| Big Five – Extraversion | Rammstedt/John (2007) | .37** | .34** | 1.06 | .23** | .36** |
| Big Five – Offenheit | Rammstedt/John (2007) | .28** | .35** | -2.43* | .21** | .29** |
| Big Five – Verträglichkeit | Rammstedt/John (2007) | .07 | .10* | -0.99 | -.02 | .04 |
| Big Five - Gewissenhaftigkeit | Rammstedt/John (2007) | .27** | .32** | -1.72 | .36** | .32** |
| Selbstwert | Collani/Herzberg (2003) | - | - | - | .55** | |
| Risikobereitschaft | in Anlehnung an SOEP | - | - | - | - | .25** |
| Arbeitszufriedenheit | in Anlehnung an SOEP | - | - | - | - | .35** |
| Zufriedenheit mit Partnerschaft | ad hoc konstruiert | - | - | - | - | .19** |
| Soziale Erwünschtheit – Übertreibung positiver Qualitäten | Kemper u.a. (2012) | - | - | - | - | .28** |
| Soziale Erwünschtheit – Untertreibung negativer Qualitäten | Kemper u.a. (2012) | - | - | - | . | .17** |
| Kristalline Intelligenz | Schipolowski u.a. (in diesem Heft) | - | - | - | - | .20** |

*Anmerkungen: * p < .05, ** p < .01. Stichprobe 1 (Welle 1): CAPI und Selbstausfüller, N = 539; Stichprobe 2: CAWI; N = 359; Stichprobe 3: CAPI, N = 1134. Lebenszufriedenheit in Anlehnung an SOEP: „Wie zufrieden sind Sie gegenwärtig, alles in allem, mit Ihrem Leben?" 11-stufige Antwortskala mit 0 = „überhaupt nicht zufrieden" bis 10 = „völlig zufrieden". Arbeitszufriedenheit in Anlehnung an SOEP 2009: „Wie zufrieden sind Sie gegenwärtig mit Ihrer Arbeit?" 11-stufige Antwortskala mit 0 = „ganz und gar unzufrieden" bis 10 „ganz und gar zufrieden". Zufriedenheit mit Partnerschaft, ad hoc konstruiert: „Wie zufrieden sind Sie gegenwärtig mit Ihrer Beziehung zu Ihrem Partner/Ihrer Partnerin?" 11-stufige Antwortskala mit 0 = „ganz und gar unzufrieden" bis 10 „ganz und gar zufrieden". Risikobereitschaft, ad hoc konstruiert: „Wie schätzen Sie sich persönlich ein: Wie risikobereit sind Sie im Allgemeinen?", 7-stufige Antwortskala: 1 = „gar nicht risikobereit" bis 7 = „sehr risikobereit".*

Tabelle 6 Validitätskoeffizienten ($r_{xy}$) der ASKU (vgl. Beierlein u.a. 2012a) mit soziodemographischen Variablen sowie sozialwissenschaftlichen Inhaltsvariablen (Stichprobe 3; $N$ = 1.134)

| Variable | Messinstrument | $r_{xy}$ |
|---|---|---|
| Soziodemographische Variablen | Destatis (2010) | |
| Alter (in Jahren) | | -.13** |
| Geschlecht (1 = männlich, 2 = weiblich) | | -.05 |
| Einkommen (in Euro) | | .24** |
| Bildung (ISCED Klassifikation) | | .19** |
| Bildung (Anzahl der Bücher im Elternhaus) | in Anlehnung an PISA (2000) | .19** |
| Einschätzung wirtschaftliche Lage in der Zukunft | ALLBUS (2010) | -.10** |
| Subjektiver Gesundheitsstatus | Andersen et al. (2007) | |
| Psychisch | | -.25** |
| Physisch | | -.29** |
| Political Efficacy (Internal) | Beierlein u.a. (2012b) | .39** |

*Anmerkungen: * p < .05, ** p < .01. Einschätzung wirtschaftliche Lage in der Zukunft (aus ALLBUS): „Was glauben Sie, wie wird Ihre eigene wirtschaftliche Lage in einem Jahr sein? Erwarten Sie, dass Ihre wirtschaftliche Lage dann: 1) wesentlich besser sein wird als heute, 2) etwas besser sein wird als heute, 3) gleichbleibt, 4) etwas schlechter sein wird, oder 5) wesentlich schlechter sein wird als heute?". Bildung (Anzahl der Bücher im Elternhaus): „Wie viele Bücher gab es in Ihrem Elternhaus bzw. in dem Haus, in dem Sie aufgewachsen sind?" 6-stufige Antwortskala: 1) „0 bis 10", 2) „11 bis 25", 3) „26 bis 50", 4) „51 bis 100", 5) „101 bis 200", 6) „201 bis 500".*

### *Kriteriumsvalidität*

Für die sozialwissenschaftliche (Umfrage-)Forschung sind insbesondere die Zusammenhänge der ASKU mit soziodemographischen Variablen und sozioökonomischen Variablen von Interesse. Mit Letzteren weist die ASKU zwar schwache, aber noch immer substanzielle Effekte auf (siehe Tabelle 6). Beispielsweise zeigen sich positive Beziehungen der Kurzskala mit den sozioökonomischen Variablen Einkommen und Bildung, letzteres gemessen über die ISCED Klassifizierung der Schulbildung und die Anzahl der Bücher im Elternhaus. Hinz, Schumacher, Albani et al. (2005) berichten in ihrer bevölkerungsrepräsentativen Studie höhere Selbstwirksamkeitserwartungen für Männer und niedrige Werte für ältere Personen. Diese Ergebnissen werden durch die vorliegende Studie nur teilweise bestätigt: Entgegen der Erwartungen zeigten sich keine Geschlechtsunterschiede. Jedoch wies die allgemeine Selbstwirksamkeitserwartung eine negative Korrelation mit dem Alter auf: Die Selbstwirksamkeitserwartung nimmt danach mit dem Alter ab.

Wie erwartet zeigte sich ein Zusammenhang der ASKU mit optimistischen, wirtschaftlichen Zukunftserwartungen (operationalisiert mittels eines Maßes aus dem ALLBUS 2010, in dem Befragte ihre eigene zukünftige wirtschaftliche Lage einschätzen sollen; vgl. Luszczynska/Gutiérrez-Dona/Schwarzer 2005). Auch mit der Political Efficacy zeigte die ASKU wie erwartet mittelstarke positive Beziehungen (vgl. Bandura 1997; Vetter 1997): Beide thematisieren Kompetenzerwartungen jedoch mit unterschiedlicher Spezifität. Die in gesundheitspsychologischen Untersuchungen postulierten Zusammenhänge zwischen der Selbstwirksamkeit und der psychischen und physischen Gesundheit lassen sich auch in den vorliegenden Daten nachweisen. Höhere Selbstwirksamkeitserwartungen gehen mit weniger Beschwerden einher.

## 5 Diskussion

Im Rahmen von drei aufeinander aufbauenden empirischen Studien wurde eine Kurzskala zur Erfassung individueller Kompetenzerwartungen konstruiert und validiert. Die „Allgemeine Selbstwirksamkeit Kurzskala" (ASKU) wurde entwickelt, um eine ökonomische Messung des psychologischen Merkmals allgemeine Selbstwirksamkeitserwartung in sozialwissenschaftlichen Untersuchungen zu ermöglichen, ohne dabei auf eine reliable und valide Erfassung verzichten zu müssen. Im Gegensatz zu der in Deutschland weit verbreiteten 10-Item-Skala von Schwarzer und Jerusalem (1999) misst die ASKU das Konstrukt Allgemeine Selbstwirksamkeit mit nur 3 Items. Die Items wurden auf der Grundlage der SWE-Skala von Jersualem und Schwarzer (1999) ausgewählt und in einem kognitiven Pretest für die anvisierte Zielgruppe, deutschsprachige, volljährige Befragte, optimiert.

Die Ergebnisse der empirischen Studien belegen, dass die ASKU trotz der erheblichen Kürzung gegenüber der etablierten SWE-Skala gute psychometrische Eigenschaften aufweist und eine reliable und valide Messung des Konstrukts erlaubt. Die Retestreliabilität der ASKU weist auf einen mittelstarken Zusammenhang zwischen zwei Messungen im 6-Wochen-Intervall hin. Dies deutet darauf hin, dass die mittels der ASKU gemessenen allgemeinen Selbstwirksamkeitserwartungen durch kontextuelle bzw. zeitliche Unterschiede beeinflusst werden können. Der situative und zeitliche Kontext der Selbstwirksamkeitserwartungen wird in den ASKU-Items nicht berücksichtigt.

Die interne Konsistenz erreichte eine für Gruppenuntersuchungen zufriedenstellende Höhe. Gegenüber der SWE zeigten sich nur geringe Reliabilitätseinbußen. Die faktorielle Validität der Skala wurde mittels konfirmatorischer Faktorenanalyse

überprüft. Die Überprüfung ergab in allen getesteten Stichproben einen sehr guten Modellfit. Zudem konnte die Eindimensionalität der ASKU empirisch belegt werden (vgl. auch Hinz/Schumacher/Albani et al. 2005). Der Vergleich der Messqualität über die drei Erhebungsmodi hinweg ergab, dass die Skala in den Erhebungsmodi CAPI und Papier-und-Bleistift auf dem höchsten getesteten Niveau (skalare Invarianz) messinvariant ist. Unterschiede zeigten sich jedoch, wenn auch die Online-Version der Skala mitberücksichtigt wurde. Unterschiedliche Ursachen können dieses Ergebnis beeinflusst haben. Die Befragungssituation im CAPI- und dem Papier-und-Bleistift-Modus ähnelte sich insofern, als dass die Befragung in Anwesenheit eines Interviewers vorgenommen wurde. Es handelt sich folglich um eine soziale Interaktionssituation. Dies war bei der Onlinebefragung nicht der Fall: Hier beantworteten die Befragten die Items selbständig vor dem Computer. Die größere Distanz zu den Untersuchungsleitern könnte die Effekte sozialer Erwünschtheit (vgl. Gravdal/Sandal 2006) reduziert haben. Ein Hinweis darauf sind die niedrigeren Mittelwerte, die in der CAWI-Bedingung im Vergleich zur CAPI-Bedingung gemessen wurde. Unterstützt wird diese Annahme zudem durch die positive Korrelation der ASKU mit dem Ausmaß sozial erwünschten Antwortverhaltens.

Darüber hinaus kann die leichte Modifizierung der Itemformulierungen einen Einfluss auf die Messäquivalenz in den drei Erhebungsmodi gehabt haben. Die Analysen der CAPI- sowie der Papier-und-Bleistift-Versionen basieren dabei auf den exakt gleichen Itemformulierungen. Allerdings fielen die Itemschwierigkeiten in der Onlineversion bei allen drei Items niedriger aus, obwohl nur zwei Items modifiziert wurden. Die Frage nach potenziellen Einflussfaktoren lässt sich auf der Basis der vorliegenden Daten nicht beantworten und sollte in zukünftigen Studien berücksichtigt werden. Neben Effekten sozialer Erwünschtheit ist zudem ein Einfluss weiterer Persönlichkeitseigenschaften auf die positive Darstellung der eigenen Person in Befragungssituation mit Interviewer denkbar. Zu diesen Persönlichkeitseigenschaften zählen zum Beispiel die Konstrukte, die seit einiger Zeit in der Forschung unter dem Begriff der „Dunklen Triade der Persönlichkeit" zusammengefasst werden (Narzissmus, Machiavellismus, Psychopathie; vgl. Rauthmann 2011).

Schließlich konnte die Validität der ASKU anhand diverser Stichproben gesichert werden. Die ASKU korreliert moderat bis hoch mit einem alternativen Maß für das Konstrukt, der SWE (Schwarzer/Jerusalem 1999), mit der internalen Kontrollüberzeugung, dem Optimismus, der Lebenszufriedenheit, den globalen Persönlichkeitsdimensionen Extraversion, Offenheit und Gewissenhaftigkeit, der allgemeinen Lebenszufriedenheit sowie dem Selbstwertgefühl. Auch zeigten sich Zusammenhänge mit der Risikobereitschaft und der kristallinen Intelligenz. Personen mit einem ausgeprägten Weltwissen fühlen sich eher kompetent, unterschied-

liche Probleme in ihrem Leben zu meistern. Negative Assoziationen zeigte die ASKU erwartungsgemäß mit dem Persönlichkeitsmerkmal Neurotizismus und mit externalen Kontrollüberzeugungen.

Die hohen Zusammenhänge der mit der ASKU gemessenen Allgemeinen Selbstwirksamkeitserwartung werfen jedoch eine wichtige Frage auf: Signalisiert die starke Überlappung mit anderen Konstrukten, dass die Selbstwirksamkeit lediglich Teil eines übergeordneten positiven Selbstkonzepts ist und sich deshalb schwer von den anderen Konstrukten unterschieden werden kann (vgl. Judge/Erez/Bono et al. 2002)? Nach dem handlungstheoretischen Partialmodell der Persönlichkeit von Krampen (2000) lässt sich jedoch die Selbstwirksamkeit (hier: Situations-Handlungs-Erwartungen bzw. Kompetenzerwartungen) zumindest theoretisch eindeutig von der Kontrollüberzeugung (hier: Handlungs-Ereignis-Erwartungen, Kompetenzerwartungen) abgrenzen. Inwiefern diese Differenzierung mittels der Erfassung über Selbstberichtsskalen auch empirisch möglich ist, ist fraglich. Es ist zu vermuten, dass Befragte bei der Beantwortung der Items gleichzeitig Informationen zu beiden Konzepten aus dem Gedächtnis abrufen. Dies lässt sich an einem Itembeispiel der IE-4-Skala (Kovaleva et al. 2012) zur internalen Kontrollüberzeugung verdeutlichen: „Wenn ich mich anstrenge, werde ich auch Erfolg haben". Die entscheidende Voraussetzung für den Erfolg zum Beispiel in einem Mathematiktest ist, dass die Person in der Lage ist, sich anzustrengen und über die nötigen Mathematikkompetenzen verfügt. Gleichzeitig muss die Bewertung der erbrachten Mathematikleistung nachvollziehbar und kontrollierbar sein. Trotz sorgfältiger Konstruktion der verschiedenen Skalen können Kontrollüberzeugungs- sowie die Selbstwirksamkeitsitems damit implizit bereits Ausprägungen des jeweils anderen Konstrukts miterfassen. Ist eine hoch differenzierte Messung für eine bestimmte Forschungsfrage von besonderer Bedeutung, empfiehlt es sich im Sinne eines Multitrait-Multimethod-Ansatzes auch alternative Erhebungsverfahren (z.B. Fremdratings, Beobachtungen, implizite Testverfahren) zur Messung der Selbstwirksamkeit mit zu erwägen. Die Entwicklung dieser alternativen Erhebungsverfahren sollte zum Gegenstand zukünftiger Studien werden.

Neben den Korrelationen mit psychologischen Variablen zeigten sich auch bedeutsame Zusammenhänge der ASKU mit soziodemographischen, sozioökonomischen und sozialwissenschaftlichen Inhaltsvariablen. Wie erwartet steht die allgemeine Selbstwirksamkeitserwartung in positivem Zusammenhang mit dem Einkommen und der Bildung; mit dem Alter zeigten sich negative Beziehungen. Jüngere, höher gebildete und Befragte mit höherem Einkommen beschreiben sich selbst eher als selbstwirksam als ältere, weniger gebildete Personen und solche mit niedrigem Einkommen. Bandura (1986) zufolge beeinflussen sich die Kogni-

tionen einer Person, ihr Verhalten und Umweltfaktoren gegenseitig („reziproker Determinismus". Vor diesem Hintergrund ist zu vermuten, dass der positive Zusammenhang zwischen Selbstwirksamkeitserwartungen und Einkommen bzw. Bildung nicht unidirektional, sondern bidirektional ist. Personen mit stärkeren Kompetenzerwartungen streben höhere Bildungsabschlüsse an, weil sie sich dafür kompetent halten. Das Erreichen des Bildungsziels wirkt sich wiederum positiv auf die Selbstwirksamkeitserwartung aus. Ein ähnlicher Zusammenhang kann in Bezug auf den (monetären) beruflichen Erfolg unterstellt werden. Im Gegensatz zu früheren Untersuchungen zeigte sich in der vorliegenden Studie kein Geschlechtseffekt auf die selbstberichteten Selbstwirksamkeitserwartungen. Diese Inkonsistenz sollte in zukünftigen Studien näher erforscht werden.

Das Design der vorliegenden Studie sowie einzelne Ergebnisse weisen auf (weitere) Grenzen der Untersuchung sowie der neu entwickelten Kurzskala hin, deren Bedeutung in zukünftigen Studien weiter eruiert bzw. berücksichtigt werden muss. Ziel war es, allgemeine Selbstwirksamkeitserwartungen zu erfassen, die unabhängig vom Kontext bzw. der Situation sind. Bandura (1997, 2006) weist jedoch zu Recht darauf hin, dass die Verhaltensvorhersage deutlich verbessert werden kann, wenn spezifischere Maße verwendet werden (siehe auch Luszczynska/Scholz/Schwarzer 2005). Eine Messung für spezifische Handlungsfelder könnte zudem die Stabilität der Messung zusätzlich erhöhen. Für die Vorhersage eines stark kontextspezifischen Verhaltens (z.B. zur Erfassung der politischen Kompetenzen) sollten deshalb kontextspezifische Instrumente bevorzugt werden, wenn die finanziellen und zeitlichen Ressourcen der Umfrage dies zulassen. Jedoch untermauert die vorliegende Studie, dass die ASKU trotz des hohen Abstraktionsgrads des gemessenen Konstrukts einen bedeutsamen Beitrag zur Erklärung sozialwissenschaftlicher Inhaltsvariablen leistet.

Wie die Prüfungen der Messinvarianz gezeigt haben, ist die Skala für die Erhebungsmodi CAPI und Papier-und-Bleistift gleichermaßen geeignet. Signifikante Unterschiede der Messergebnisse fanden sich jedoch in der Online-Studie. Demzufolge konnte die Messäquivalenz nur für zwei der drei Erhebungsmodi abgesichert werden. Die Ergebnisse deuten darauf hin, dass die Itemschwierigkeiten in der Online-Version niedriger lagen als in den beiden anderen Erhebungsmodi. Dieses Ergebnis muss bei der Durchführung zukünftiger Studien mit der ASKU beachtet werden. Jedoch gelten die günstigen psychometrischen Kennwerte der Skala auch im Online-Erhebungsmodus.

Insgesamt legen die berichteten Belege zur Güte der ASKU nahe, dass die Skala eine reliable, valide und ökonomische Erfassung von subjektiven Kompetenzerwartungen in der sozialwissenschaftlichen Forschung erlaubt. Aufgrund ihrer

hohen Ökonomie ist sie insbesondere für Untersuchungen geeignet, die starken zeitlichen oder monetären Restriktionen unterliegen. Dies gilt zum Beispiel für sozialwissenschaftliche Bevölkerungsumfragen und Onlinestudien.

## Literatur


Akhtar, S., G. Saba, G. und A. Adnan, in Druck: Self-efficacy and optimism as predictors of organizational commitment among bank employees. International Journal of Research Studies in Psychology.

Andersen, H. H., A. Mühlbacher und M. Nübling, M., 2007: Die SOEP-Version des SF 12 als Instrument gesundheitsökonomischer Analysen. Data Documentation 6. Berlin: Deutsches Institut für Wirtschaftsforschung (DIW). http://www.diw.de/documents/publikationen/73/diw_01.c.56544.de/diw_sp0006.pdf

Bandura, A., 1977: Self-efficacy: Toward a unifying theory of behavioral change. Psychological Review 84: 191215.

Bandura, A., 1982: Self-efficacy mechanism in human agency. *American Psychologist, 33*, 344-358.

Bandura, A., 1986: Social Foundations of Thought and Action. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A., 1993: Perceived self-efficacy in cognitive development and functioning. Educational Psychologist 28: 117-148.

Bandura, A., 1997: Self-efficacy: The exercise of control. New York: Freeman.

Bandura, A., 2006: Guide for constructing self-efficacy scales. S. 307-337 in F. Pajares u. T. Urdan (Hg.): Self-efficacy beliefs of adolescents (Band 5). Greenwich, CT: Information Age Publishing.

Beauducel, A. und W. W. Wittmann, 2005: Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. Structural Equation Modeling 12: 41–75.

Beierlein, C., C. J. Kemper, A. Kovaleva und B. Rammstedt, 2012a: Ein Messinstrument zur Erfassung subjektiver Kompetenzerwartungen: Allgemeine Selbstwirksamkeit Kurzskala (ASKU). GESIS-Working Papers 2012|17. Köln: GESIS. http://www.gesis.org/fileadmin/kurzskalen/working_papers/ASKU_Workingpaper.pdf

Beierlein, C., C. J. Kemper, A. Kovaleva und B. Rammstedt, 2012b: Ein Messinstrument zur Erfassung politischer Kompetenz- und Einflusserwartungen: Political Efficacy Kurzskala (PEKS). GESIS-Working Papers 2012|18. Köln: GESIS. http://www.gesis.org/fileadmin/kurzskalen/working_papers/PEKS_Workingpaper.pdf

Breakwell, G. M., 2007: The psychology of risk. Cambridge: Cambridge University Press.

Brown, T. A., 2006: Confirmatory factor analysis for applied research. New York: Guilford.

Browne, M. W. und R.Cudeck, R., 1993: Alternative ways of assessing model fit. Sociological Methods & Research, 21: 230–258.

Bühner, M., 2011: Einführung in die Test- und Fragebogenkonstruktion (3. aktualisierte und erw. Auflage). München: Pearson-Education.

Byrne, B., 2009: *Structural equating modelling with AMOS: Basic concepts, applications, and programming* (2. Auflage). New York, NY: Taylor & Francis Group.

Byrne, B. M., R. J. Shavelson und B. Muthén, 1989: Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin, 105: 456-466.

Caprara, G. V., 2002: Personality psychology: Filling the gap between basic processes and molar functioning. S. 201-224 in C. von Hofsten und L. Bakman (Hg.): Psychology at the turn of the Millennium: Volume 2. Social, developmental and clinical perspectives. Hove, UK: Psychology Press.

Caprara, G. V., G. Alessandri und N. Eisenberg, 2012: Prosociality: The contribution of traits, values, and self-efficacy beliefs. Journal of Personality and Social Psychology 102: 1289-1303.

Caprara, G. V., R. Fida, M. Vecchione, G. Del Bove, G. M. Vecchio, C. Barbaranelli und A. Bandura, 2008: Longitudinal analysis of the role of perceived self-efficacy for self-regulated learning in academic continuance and achievement. Journal of Educational Psychology 100: 525-534.

Caprara, G. V. und P. Steca, 2005: Affective and Social Self-Regulatory Efficacy Beliefs as Determinants of Positive Thinking and Happiness. European Psychologist 10: 275-286.

Caprara, G. V., M. Vecchione, G. Alessandri, M. Gerbino und C. Barbaranelli, 2011: The contribution of personality traits and self-efficacy beliefs in academic achievement: A longitudinal study. British Journal of Educational Psychology 81: 78-96.

Cheung, G. W. und R. B. Rensvold, 2002: Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling 9: 233-255.

Cohen, J., 1992: *A power primer*. Psychological Bulletin 112: 155-159.

Collani, G. von und P. Y. Herzberg, 2003. Eine revidierte Fassung der deutschsprachigen Skala zum Selbstwertgefühl von Rosenberg. Zeitschrift für Differentielle und Diagnostische Psychologie, 24: 3-7.

Destatis, 2010: Demographische Standards (5. und erweiterte Auflage). Wiesbaden: Statistisches Bundesamt. https://www.destatis.de/DE/Methoden/DemografischeRegionale Standards/DemografischeStandardsInfo.html?nn=173768

Diener, E., R. A. Emmons, R. J. Larsen und S. Griffin, 1985: The satisfaction with life scale. Journal of Personality Assessment 49: 7175.

Faulbaum, F., P. Prüfer und M. Rexroth, 2009: *Was ist eine gute Frage?* – Die systematische Evaluation der Fragenqualität. Wiesbaden: VS Verlag für Sozialwissenschaften.

Glaesmer, H., J. Hoyer, J. Klotsche und P. Y. Herzberg, 2008: Die deutsche Version des Life-Orientation-Tests (LOT-R) zum dispositionellen Optimismus und Pessimismus. Zeitschrift für Gesundheitspsychologie 16: 2631.

Gravdal, L. und G. M. Sandal, 2006: The two factor model of social desirability. Relations to coping and defense, and implications for health. Personality and Individual Differences 40: 1051-1061.

Gwalthney, C. J., J. Metrik, C. W. Kahler und S. Shiffman, 2009: Self-efficacy and smoking cessation: a meta-analysis. Psychology of Addictive Behaviors 23: 56-66.

Heuven, E., A. B. Bakker, W. B. Schaufeli und N. Huisman 2006: The role of self-efficacy in performing emotion work. Journal of Vocational Behavior 69: 222-235.

Hinz, A., J. Schumacher, C. Albani, G. Schmid und E. Brähler, 2006: Bevölkerungsrepräsentative Normierung der Skala zur Allgemeinen Selbstwirksamkeitserwartung. Diagnostica 52: 2632.

Hu, L. und P. M. Bentler, 1999: Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling 6: 1–55.

Ilgen, M., J. McKellar und Q. Tiet, 2005: Abstinence self-efficacy and abstinence 1 year after substance use disorder treatment. Journal of Consulting and Clinical Psychology 73: 1175-1180.

Jakoby, N. und R. Jacob, (1999): Messung von internen und externen Kontrollüberzeugungen. ZUMA-Nachrichten 45: 6171.

Jerusalem, M. und R. Schwarzer, 1986: Selbstwirksamkeit. S. 15-28 in R. Schwarzer (Hg.): Skalen zur Befindlichkeit und Persönlichkeit. Berlin: Institut für Psychologie, Freie Universität Berlin.

Jerusalem, M. und R. Schwarzer, 1999: Skala Allgemeine Selbstwirksamkeitserwartung. S. 16-17 in: R. Schwarzer und M. Jerusalem (Hg.): Skalen zur Erfassung von Lehrer- und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen. Berlin: Freie Universität Berlin.

Jöreskog, K. G., 1969: A general approach to confirmatory maximum likelihood factor analysis. Psychometrika 34: 183202.

Judge, T. A. und J. E. Bono, 2001: Relationship of core self-evaluations traits - self-esteem, generalized self-efficacy, locus of control, and emotional stability - with job satisfaction and job performance: A meta-analysis. Journal of Applied Psychology 86: 80-92.

Judge, T. A., A. Erez, J. E. Bono und C. J. Thoresen, 2002: Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? Journal of Personality and Social Psychology 83: 693-710.

Judge, T. A. und R. Ilies, 2002: Relationship of personality to performance motivation: A meta-analysis. Journal of Applied Psychology 87: 797-807.

Kemper, C. J., C. Beierlein, D. Bensch, A. Kovaleva und B. Rammstedt, 2012: Eine Kurzskala zur Erfassung des Gamma-Faktors sozial erwünschten Antwortverhaltens: Die Kurzskala Soziale Erwünschtheit-Gamma (KSE-G). GESIS-Working Papers 2012|25). Köln: GESIS.

Kemper, C. J., C. Beierlein, A. Kovaleva und B. Rammstedt, in Druck: Entwicklung und Validierung einer ultrakurzen Operationalisierung des Konstrukts Optimismus-Pessimismus. Diagnostica.

Kline, R. B., 2011: Principles and practice of structural equation modeling (3. Auflage). New York, London: The Guilford Press.

Kovaleva, A., C. Beierlein, C. J. Kemper und B. Rammstedt, 2012: Eine Kurzskala zur Messung von Kontrollüberzeugung: Die Skala Internale-Externale-Kontrollüberzeugung-4 (IE-4). GESIS-Working Papers 2012|19. Köln: GESIS. http://www.gesis.org/fileadmin/kurzskalen/working_papers/IE4_Workingpaper.pdf

Krampen, G., 2000: Handlungstheoretische Persönlichkeitspsychologie. Konzeptuelle und empirische Beiträge zur Konstrukterhellung (2., erg. Auflage). Göttingen: Hogrefe.

Krohne, H. W. und M. Hock, 2007: Psychologische Diagnostik: Grundlagen und Anwendungsfelder. Stuttgart: Kohlhammer.

Leganger, A., P. Kraft und E. Roysamb, 2000: Perceived self-efficacy in health: Conceptualisation, measurement, correlates. Psychology & Health 15: 51-69.

Luszczynska, A., C. C. Benight und R. Cieslak, 2009: Self-efficacy and health-related outcomes of collective trauma: A systematic review. European Psychologist 14: 49-60.

Luszczynska, A., B. Gutiérrez-Doña und R. Schwarzer, 2005: General self-efficacy in various domains of human functioning: Evidence from five countries. International Journal of Psychology 40: 8089.

Luszczynska, A., U. Scholz und R. Schwarzer, 2005: The general self-efficacy scale: Multicultural validation studies. The Journal of Psychology 139: 439-457.

Magaletta, P. R. und J. M. Oliver, 1999: The hope construct, will, and ways: their relations with self-efficacy, optimism, and general well-being. Journal of Clinical Psychology 55: 539-551.

McCrae, R. R. und P. T. Costa, 1987: Validation of the five-factor model of personality across instruments and observers. Journal of Personality and Social Psychology 52: 81–90.

McDonald, R. P., 1999: Test theory: A unified treatment. Mahwah: Erlbaum.

Muthén, L. K. und B. O. Muthén, 1998-2010: Mplus User's Guide (Sixth Edition). Los Angeles, CA: Muthén und Muthén.

Pajares, F., 1997: Current directions in self-efficacy research. S. 1-49 in: M. Maehr und P. R. Pintrich (Hg.): Advances in motivation and achievement (Vol. 10). Greenwich, CT: JAI Press.

Pajares, F., J. Hartley und G. Valiante, 2001: Response format in writing self-efficacy assessment: Greater discrimination increases prediction. Measurement and Evaluation in Counseling and Development 33: 214-221.

Pells, J. J., R. A. Shelby, F. J. Keefe, K. E. Dixon, J. A. Blumenthal, L. Lacaille, J. M. Tucker, D. Schmitt, D. S. Caldwell und V. B. Kraus, 2008: Arthritis self-efficacy and self-efficacy for resisting eating: relationships to pain, disability, and eating behavior in overweight and obese individuals with osteoarthritic knee pain. Pain 136: 340-347.

Prüfer, P. und M. Rexroth, 2000: Zwei-Phasen-Pretesting. ZUMA-Arbeitsbericht 8: 128. http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/zuma_arbeitsberichte/00_08.pdf

Rammstedt, B. und O. P. John, 2007: Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of Research in Personality 41: 203212.

Rauthmann, J. F. 2011: Acquisitive or protective self-presentation of dark personalities? Associations among the Dark Triad and self-monitoring. Personality and Individual Differences 51: 502-508.

Rohrmann, B., 1978: Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. Zeitschrift für Sozialpsychologie 9: 222–245.

Rosenberg, M., 1989: Society and the adolescent self-image. Middletown: Wesleyan University Press.

Rotter, J. B., 1966: Generalized expectancies for internal and external control of reinforcement. Psychological Monographs 80.

Scherbaum, C. A., Y. Cohen-Charash und M. J. Kern, 2006: Measuring general self-efficacy. A comparison of three measures using item response theory. Educational and Psychological Measurement 66: 1047-1063.

Schipolowski, S. et al., 2013: BEFKI GC-K: Eine Kurzskala zur Messung kristalliner Intelligenz. mda Jg. 7 (2), 155-184.

Scholz, U., B. Gutiérrez-Doña, S. Sud und R. Schwarzer, 2002: Is general self-efficacy a universal construct? Psychometric findings from 25 countries. European Journal of Psychological Assessment 18: 242-251.

Schunk, D. H., 1991: Self-efficacy and academic motivation. Educational Psychologist 26: 207231.

Schwarzer, R., 1994: Optimistische Kompetenzerwartung: Zur Erfassung einer personalen Bewältigungsressource. Diagnostica 40: 105123.

Schwarzer, R. und A. Born, 1997: Optimistic self-beliefs: Assessment of general perceived self-efficacy in thirteen cultures. World Psychology 3: 177-190.

Schwarzer, R., J. Bäßler, P. Kwiatek, K. Schröder und J. X. Zhang, 1997: The assessment of optimistic self-beliefs: Comparison of the German, Spanish, and Chinese versions of the General Self-Efficacy Scale. Applied Psychology: An International Review 46: 69-88.

Schweizer, K., 2011. On the changing role of Cronbach α in the evaluation of the quality of a measure. European Journal of Psychological Assessment 27: 143144.

Siegrist, J., D. Starke, T. Chandola, I. Godin, M. Marmot, I. Niedhammer und R. Peter, 2004: The measurement of effort-reward imbalance at work: European comparisons. Social science & medicine 58: 14831499.

Sirtharthan, T., R. F. S. Job, D. J. Kavanagh, G. Sirtharthan und M. Hough, 2003: Development of a controlled drinking self-efficacy scale and appraising its relation to alcohol dependence. Journal of Clinical Psychology 59: 351-362.

Skaalvik, E. M. und S. Skaalvik, 2007. Dimensions of teacher self-efficacy and relations with strain factors, perceived collective teacher efficacy, and teacher burnout. Journal of Educational Psychology 99: 611-625.

Turner, J. A., M. Ersek und C. A. Kemp, 2005: Self-efficacy for managing pain is associated with disability, depression, and pain coping among retirement community residents with chronic pain. Journal of Pain 6: 471-479.

Vandenberg, R. J. und C. E. Lance, 2000: A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods 3: 4-69.

Vecchione, M. und G. V. Caprara, 2009: Personality determinants of political participation: the contribution of traits and self-efficacy beliefs. Personality and Individual Differences 46: 487-492.

Vetter, A., 1997: Political Efficacy: Alte und neue Messmodelle im Vergleich. Kölner Zeitschrift für Soziologie und Sozialpsychologie 49: 53-73.

Williams, T. und K. Williams, 2010: Self-efficacy and performance in mathematics: Reciprocal determinism in 33 nations. Journal of Educational Psychology 102: 453-466.

Zimmerman, B. J. und T. J. Cleary, 2006: Adolescents' development of personal agency. The role of self-efficacy beliefs and self-regulatory skill. S. 45-70 in: F. Pajares und T. Urdan (Hg.): Self-efficacy beliefs of adolescents. Greenwich, CT: Information Age Publishing.

Zuffianò, A., G. Alessandri, M. Gerbino, B. P. Luengo Kanacri, L. Di Giunta, M. Milioni und G.V. Caprara (in Druck). Academic Achievement: The unique contribution of self-efficacy beliefs in self-regulated learning beyond intelligence, personality traits, and self-esteem. Learning and Individual Differences.

Anschrift der Autorin

Constanze Beierlein
GESIS – Leibniz-Institut für Sozialwissenschaften
B2,1
68159 Mannheim.
E-Mail: constanze.beierlein@gesis.org

Ko-Autor/-innen

Christoph J. Kemper
Institut für Medizinische und
Pharmazeutische Prüfungsfragen (IMPP), Mainz

Anastassiya Kovaleva
Institut für Biologie, Universität Bielefeld

Beatrice Rammstedt
GESIS – Leibniz-Institut für Sozialwissenschaften
Mannheim

# Appendix

## Allgemeine Selbstwirksamkeit Kurzskala (ASKU)

Die folgenden Aussagen können mehr oder weniger auf Sie zutreffen. Bitte geben Sie bei jeder Aussage an, inwieweit diese auf Sie persönlich zutrifft.

| | trifft gar nicht zu | trifft wenig zu | trifft etwas zu | trifft ziemlich zu | trifft voll und ganz zu |
|---|---|---|---|---|---|
| (1) In schwierigen Situationen kann ich mich auf meine Fähigkeiten verlassen. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| (2) Die meisten Probleme kann ich aus eigener Kraft gut meistern. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |
| (3) Auch anstrengende und komplizierte Aufgaben kann ich in der Regel gut lösen. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 |

methoden, daten, analysen · 2013, Jg. 7(2), S. 279-310 DOI: 10.12758/mda.2013.015

# Vier Kurzskalen zur Messung des Persönlichkeitsmerkmals „Sensibilität für Ungerechtigkeit“

# Four Short Scales for Measuring the Personality Trait of “Justice Sensitivity”

*Constanze Beierlein, Anna Baumert, Manfred Schmitt, Christoph J. Kemper und Beatrice Rammstedt*

*Zusammenfassung*
Menschen unterscheiden sich systematisch darin, wie leicht sie Ungerechtigkeit wahrnehmen und wie stark sie darauf reagieren. Das Konstrukt der Ungerechtigkeitssensibilität spiegelt diese stabilen Persönlichkeitsunterschiede wider. Empirische Studien zeigen, dass die Ungerechtigkeitssensibilität mit Variablen aus den Bereichen Arbeit, Gesundheit, Politik und soziale Beziehungen korreliert ist. Schmitt, Baumert, Gollwitzer und Maes (2010) legten ein erstes Messinstrument zur Erfassung der vier Facetten der Ungerechtigkeitssensibilität vor. Darauf aufbauend entwickelten Baumert, Beierlein, Schmitt, Kemper, Kovaleva, Liebig und Rammstedt (in Druck) vier für die Surveyforschung geeignete Kurzskalen mit je zwei Items. In dem vorliegenden Artikel wird der Konstruktionsprozess der Kurzskalen dargestellt. Validierungsergebnisse werden auf der Basis empirischer Daten aus einer quoten- und einer bevölkerungsrepräsentativen Stichprobe berichtet. Die Ergebnisse belegen die zufriedenstellende Reliabilität und Validität der Kurzskalen. Die Kurzskalen

*Abstract*
People differ systematically in how easily they perceive injustice and in how strongly they respond to it. The construct of justice sensitivity mirrors these stable personality differences. Empirical studies have demonstrated that justice sensitivity is correlated with variables from the field of work, health, politics, and social relations. Schmitt, Baumert, Gollwitzer, and Maes (2010) developed a first instrument measuring the four facets of justice sensitivity. Based on this instrument, Baumert, Beierlein, Schmitt, Kemper, Kovaleva, Liebig, and Rammstedt (in press) constructed four short scales with two items each which are particularly suitablefor survey research. In the present paper, we explain how the short scales were constructed. Empirical results on the validity of the scales will be reported on the basis of a quota sample and a population representative sample. The findings corroborate the satisfactory reliability and validity of the short scales. The factorial structure was appropriately replicated by the short scales. Furthermore, the

konnten die faktorielle Struktur der Ungerechtigkeitssensibilität angemessen replizieren. Überdies weisen die Kurzskalen die theoretisch erwarteten Beziehungen zu Kriteriumsvariablen auf (z.B. Effort-Reward-Imbalance, psychische Gesundheit, Political Efficacy, Delinquenz).

short scales showed the theoretically expected relations to criterion variables such as effort-reward-imbalance, mental health, political efficacy, and delinquency.

# 1 Einleitung[1]

Fragen der Gerechtigkeit sind in den letzten Jahren wieder stärker in den Mittelpunkt der gesellschaftlichen Aufmerksamkeit gerückt: Insbesondere die Auswirkungen der Finanzkrise von 2008 und die Protestbewegungen im Rahmen des „Arabischen Frühlings" ab Dezember 2010 werden aktuell mit Fragen der Verteilungsgerechtigkeit in Verbindung gebracht (Kraushaar 2012; Noll/Weick 2012). Die Diskussion um Verteilungsfragen in der Gesellschaft wurde auch durch die Veröffentlichung des Bandes „The Spirit Level. Why more equal societies almost always do better" der britischen Epidemiologen Richard Wilkinson und Kate Pickett (2009) weiter verstärkt.

Die empirische Gerechtigkeitsforschung ist zurzeit eines der bedeutendsten interdisziplinären sozialwissenschaftlichen Forschungsgebiete: In der Soziologie, in der Politikwissenschaft als auch in der Ökonomie spielen gerechtigkeitsbezogene Fragen eine herausgehobene Rolle (Fetchenhauer/Goldschmidt/Hradil/Liebig 2010). Beforscht werden unter anderem Fragen der Einkommens- und Steuergerechtigkeit, der Verfahrensgerechtigkeit in organisationalen Prozessen, der Beziehung zwischen Gerechtigkeitserleben und der Legitimation politischer Systeme, Gerechtigkeit in internationalen Beziehungen sowie subjektive Gerechtigkeitsbeurteilungen (für einen Überblick siehe z.B. Liebig/Lengfeld 2002; Liebig/Lengfeld/Mau 2004).

Auch die psychologische Gerechtigkeitsforschung leistete zur Entwicklung dieses Forschungszweigs einen zentralen Beitrag. Liebig (2004) weist darauf hin, dass die empirische Gerechtigkeitsforschung bis in die 1990er Jahre hinein insbesondere durch Erkenntnisse aus der psychologischen Gerechtigkeitsforschung bestimmt war. Viele theoretische Modelle und empirische Studien basierten dabei auf sozialpsychologischen Annahmen. Seit Mitte der 1990er Jahre werden in der psychologischen Gerechtigkeitsforschung zunehmend Persönlichkeitsunterschiede im Erleben von und in Reaktionen auf Ungerechtigkeit erforscht (vgl. Schmitt

1 Die Erstautorin bedankt sich bei Eldad Davidov sowie den anonymen Gutachtern für methodische Hinweise.

1996). Das Konstrukt der Ungerechtigkeitssensibilität spiegelt diese dispositionellen Unterschiede wider: Danach unterscheiden sich Menschen systematisch darin, wie leicht sie Ungerechtigkeit wahrnehmen und wie stark sie darauf reagieren (Schmitt/Baumert/Fetchenhauer/Gollwitzer/Rothmund/Schlösser 2009). Schmitt, Gollwitzer, Maes und Arbach (2005) konnten zeigen, dass diese Unterschiede zeitlich stabil sind und sich über ungerechte Situationen hinweg generalisieren lassen.

Ungerechtigkeit kann aus vier Perspektiven wahrgenommen werden: aus der Opfer-, der Beobachter-, der Nutznießer- und der Täterperspektive (Schmitt et al. 2009). Den vier Perspektiven ist gemeinsam, dass sich in ihnen eine allgemeine Sorge um Gerechtigkeit ausdrückt. Es lassen sich jedoch auch Differenzen zwischen den vier Perspektiven aus der Theorie ableiten: Demzufolge sind diese mit unterschiedlichen Emotionen und Verhaltenstendenzen verknüpft (Schmitt et al. 2009). Bei der Beobachter-, der Nutznießer- und der Tätersensibilität kommt der Wunsch nach Gerechtigkeit für andere und das Gefühl sozialer Verantwortung zum Ausdruck. Dagegen vereint die Opfersensibilität eine Mischung aus selbstbezogenen und gerechtigkeitsbezogenen Sorgen (Gollwitzer/Schmitt/Schalke/Maes/Baer 2005). Opfersensible sind besonders empfänglich für Ungerechtigkeit, die sie selbst betrifft. Sie reagieren auf dieses Erlebnis mit Ärger. Beobachtersensible reagieren mit starker Empörung, wenn sie Zeuge bzw. Zeugin von Ungerechtigkeit werden. Nutznießersensible sind besonders besorgt um die Gerechtigkeit, wenn sie selbst von einer Ungerechtigkeit profitieren könnten. Tätersensible zeigen starke kognitive und emotionale Reaktionen, wenn sie selbst zum Täter bzw. zur Täterin einer ungerechten Handlung werden könnten. Nutznießer- und Tätersensible reagieren in erster Linie mit Schuldgefühlen auf die geschilderten Ungerechtigkeitssituationen. Die Nutznießersensibilität geht dabei auf das Konzept der „Existenziellen Schuld" zurück. Danach können Personen Profiteure von Ungerechtigkeit sein, müssen diese aber nicht selbst verursacht haben (Montada/Dalbert/Reichle/Schmitt 1985). Die vier Perspektiven der Ungerechtigkeitssensibilität konnten in Studien empirisch voneinander getrennt werden (z.B. Schmitt et al. 2010).

Mehrere Studien belegen, dass die vier Perspektiven der Ungerechtigkeitssensibilität theoretisch erwartete Beziehungen mit psychologischen, soziologischen und politikwissenschaftlichen Variablen aufweisen. Die Ungerechtigkeitssensibilität leistet dabei einen Beitrag zur Vorhersage sozialwissenschaftlicher Inhaltsvariablen, der über denjenigen der Big-Five-Persönlichkeitsdimensionen hinausgeht (vgl. Schmitt et al. 2010). Das macht sie insbesondere für sozialwissenschaftliche Umfragen interessant. Zusammenhänge konnten nachgewiesen werden mit Variablen aus den Bereichen soziale Beziehungen im Privatleben, Gesundheit, Arbeitsle-

ben und politische Einstellungen. Der folgende Abschnitt soll einen Überblick über die Ergebnisse aus einigen zentralen Studien bieten.

Schmitt und Kollegen (2010) zufolge gehen die Beobachter-, die Nutznießer- und die Tätersensibilität mit Facetten der Big-Five-Persönlichkeitsdimension Verträglichkeit einher (z.B. Bescheidenheit). Studien belegen zudem, dass diese drei Perspektiven der Ungerechtigkeitssensibilität positiv mit prosozialem Verhalten (z.B. Solidarität mit Benachteiligten, Altruismus, kooperativem Verhalten) korrelieren (Gollwitzer/Rothmund/Pfeiffer/Ensenbach 2009; Gollwitzer et al. 2005). Im Gegensatz dazu zeigt die Opfersensibilität negative Beziehungen zu prosozialem Verhalten. Die Ergebnisse von Schmitt und Kollegen (2005) liefern hierfür weitere Hinweise: Eine hohe Opfersensibilität war in ihrer Studie mit antisozialen Tendenzen wie z.B. Machiavellismus, Paranoia, Neurotizismus und Eifersucht assoziiert. Opfersensible sehen die Welt demzufolge als einen gefährlichen Ort an, an dem die Gefahr lauert, ausgebeutet und übervorteilt zu werden.

Opfersensible zeigen auch weniger Vertrauen zu anderen Menschen und haben stärker ihre eigenen Interessen im Blick (Schmitt et al. 2005; Schmitt et al. 2009). Diese Verhaltensweisen können sich in sozialen Beziehungen (z.B. in romantischen Partnerschaften) negativ auswirken. So konnten Gerlach, Allemand, Agroskin und Denissen (2012) zeigen, dass Opfersensible weniger bereit sind, ihrem Partner oder ihrer Partnerin bei einem Konflikt zu vergeben. Dieses Verhalten zeigten die Opfersensiblen auch dann, wenn sich der Partner/die Partnerin um Wiedergutmachung der begangenen Ungerechtigkeit bemühte. Fetchenhauer und Huang (2003) berichten, dass Opfersensible in einem Strategiespiel eher Entscheidungen zu ihren eigenen Gunsten trafen. Hoch Nutznießersensible Personen entschieden sich dagegen häufiger für Angebote mit gleicher Aufteilung für sich selbst und andere als Personen mit niedriger Ausprägung in diesem Persönlichkeitsmerkmal.

Baumert, Beierlein, Schmitt, Kemper, Kovaleva, Liebig und Rammstedt (in Druck) weisen auf die Bedeutung der Ungerechtigkeitssensibilität als wichtigen Risikofaktor für das *physische und psychische Wohlbefinden* hin. Schmitt und Dörfel (1999) fanden in einer Studie an Arbeitnehmerinnen und Arbeitnehmern heraus, dass die Opfersensibilität den Effekt von prozeduraler Gerechtigkeit auf das psychosomatische Wohlbefinden erhöht. Je höher diese Facette der Ungerechtigkeitssensibilität, desto stärker war der Zusammenhang zwischen dem Ausmaß an erlebter prozeduraler Ungerechtigkeit und der Anzahl der Arbeitstage, an denen die Person über psychosomatischen Beschwerden klagte.

Darüber hinaus steht die Ungerechtigkeitssensibilität in Zusammenhang mit weiteren Phänomenen des Erlebens und Verhaltens in der *Arbeitswelt*, z.B. Loyalität gegenüber dem Arbeitgeber oder Racheintentionen infolge von Kündigung des

Arbeitsplatzes. Schmitt, Rebele, Bennecke und Förster (2008) konnten zeigen, dass Einstellungen und Verhaltenstendenzen von Gekündigten gegenüber ihrem ehemaligen Arbeitgeber Beziehungen zur Ungerechtigkeitssensibilität aufweisen: Je höher in der Studie die Ungerechtigkeitssensibilität einer Person war, desto geringer waren nach der Kündigung die positiven Einstellungen der Person gegenüber dem früheren Arbeitgeber und desto ausgeprägter der Wunsch, dem Arbeitgeber die Kündigung heimzuzahlen. Hessler, Pretsch, Hillert und Schmitt (in Vorbereitung) berichten, dass Lehrkräfte, die unter Burnout leiden, im Vergleich zu gesunden Lehrkräften höhere Ausprägungen in den Dimensionen Opfer- und Beobachtersensibilität aufweisen. Vermittelt waren diese Effekte über die Wahrnehmung eines Ungleichgewichts zwischen den subjektiv erbrachten beruflichen Leistungen und den wahrgenommenen Gegenleistungen. Dieses Ungleichgewicht wird als Gratifikationskrise bezeichnet (Siegrist/Starke/Chandola/Godin/Marmot/Niedhammer/Peter 2004; Siegrist/Wege/Pühlhofer/Wahrendorf 2008).

Auch mit Variablen aus dem *politischen Bereich* werden in der Literatur substanzielle Beziehungen der vier Perspektiven der Ungerechtigkeitssensibilität berichtet: In einer Studie von Traut-Mattausch, Guter, Zanna, Jonas und Frey (2011) zeigten hoch Opfersensible einen größeren Widerstand gegen eine politische Reform als niedrig Opfersensible. Die Autoren erklären diese Beobachtung damit, dass Opfersensible die Situation als eine illegitime Einschränkung ihrer eigenen Freiheiten interpretierten. Auf diese Interpretation hin entwickelte sich bei ihnen Reaktanz, die schließlich zur Ablehnung der Reform führte.

Auf der Basis von Bevölkerungsumfragen ließen sich zudem statistisch signifikante Unterschiede zwischen Mitgliedern unterschiedlicher soziodemographischer Gruppen zeigen. So fanden Schmitt und Kollegen (2010) heraus, dass sich Frauen bzw. Ostdeutsche insgesamt eine höhere Ungerechtigkeitssensibilität zuschrieben als Männer bzw. Westdeutsche. Im Hinblick auf das Alter zeigte sich, dass die Opfersensibilität der Befragten mit zunehmendem Alter abnahm. Eine höhere Bildung der Befragten ging mit höheren Ausprägungen der Tätersensibilität einher.

## 2 Erfassung der Ungerechtigkeitssensibilität

Schmitt, Neumann und Montada publizierten 1995 erste Selbstberichtsskalen, um die Sensibilität für widerfahrende Ungerechtigkeit anhand von vier Indikatoren zu messen: Häufigkeit erinnerter Ungerechtigkeitserlebnisse, Stärke der emotionalen Reaktion auf erfahrene Ungerechtigkeit, Penetranz von Gedanken an eine wider-

fahrene Ungerechtigkeit sowie Bestrafungs- und Vergeltungswünsche gegenüber dem Täter. Diese ersten Skalen beschränkten sich auf die Messung von Opfersensibilität. In einem zweiten Entwicklungsschritt wurde aus dieser ersten Langfassung eine kürzere Fassung erstellt, die 10 Items umfasste. Die Items beinhalten drei Indikatoren, (1) die spezifische emotionale Reaktion auf widerfahrene Ungerechtigkeit (Beispiel: Es ärgert mich, wenn es anderen unverdient besser geht als mir.), (2) die unspezifische emotionale Belastung durch widerfahrene Ungerechtigkeit (Beispiel: Es macht mir zu schaffen, wenn ich mich für Dinge abrackern muss, die anderen in den Schoß fallen.) sowie (3) die Penetranz von Gedanken an eine widerfahrene Ungerechtigkeit (Beispiel: Wenn andere ohne Grund freundlicher behandelt werden als ich, geht mir das lange durch den Kopf.). Die von Schmitt und Kollegen (1995) vorgeschlagenen Indikatoren der Häufigkeit erinnerter Ungerechtigkeitserlebnisse und der Bestrafungs- und Vergeltungswünsche wurden nicht einbezogen, weil diese Indikatoren eine geringere konvergente Validität aufwiesen als die anderen Indikatoren. Gleichzeitig wurden 10-Item-Skalen zur Messung von Beobachtersensibilität und von Nutznießersensibilität entwickelt (Schmitt et al. 2005). Diesen drei Skalen wurde später eine 10-Item-Skala zur Messung von Tätersensibilität hinzugefügt (Schmitt et al. 2010). Alle vier 10-Item-Skalen verwenden die oben genannten Indikatoren, wobei die spezifische emotionale Reaktion über die Skalen variiert (Ärger bei Opfersensibilität, Empörung bei Beobachtersensibilität, Schuldgefühle bei Nutznießer- und Tätersensibilität). Aufbauend auf diesen vier 10-Item-Skalen entwickelten Baumert und Kollegen (in Druck) vier Kurzskalen zur Messung der Ungerechtigkeitssensibilität (USS-8; Ungerechtigkeitssensibilität-Kurzskalen). Die Kurzskalen sollen den besonderen Anforderungen an Messinstrumente in sozialwissenschaftlichen Umfragen Rechnung tragen, indem sie mit wenigen Items das interessierende Konstrukt reliabel und valide messen. Auf diese Weise schonen sie zeitliche und finanzielle Ressourcen. Die Kurzskalen haben darüber hinaus den Anspruch, das Merkmal der Ungerechtigkeitssensibilität in Stichproben von Befragten, die sich im Hinblick auf Bildungsgrad, das Alter, das Geschlecht sowie den Wohnort unterscheiden (vgl. Schmitt et al. 2010), in gleicher Messqualität erfassen zu können. Vor diesem Hintergrund werden im Rahmen des vorliegenden Beitrags die folgenden Fragestellungen untersucht:

Zunächst soll überprüft werden, ob die Kurzskalen im Hinblick auf ihre Reliabilität und Validität zufriedenstellende Kennwerte aufweisen. Die psychometrischen Eigenschaften der Kurzskalen werden anhand einer bevölkerungsrepräsentativen Stichprobe getestet.

Darüber hinaus wird untersucht, ob das neue Messinstrument für Personen mit unterschiedlichen kognitiven Fähigkeiten bzw. Bildungsgraden, unterschiedlichem Alter, Geschlecht und Wohnort in Deutschland gleichermaßen geeignet ist.

Wie oben dargestellt werden Unterschiede zwischen soziodemographischen Gruppen in den einzelnen Komponenten der Ungerechtigkeitssensibilität berichtet (z.B. Schmitt et al. 2010). Die Ergebnisse dieser Studien basieren jedoch auf Erhebungen mit den Langskalen. In den vorliegenden Studien wird untersucht, ob sich die in der Literatur berichteten soziodemographischen Unterschiede auch mit den neuen Kurzskalen abbilden lassen.

Darüber hinaus soll im Rahmen der Skalenvalidierung erforscht werden, welche Beziehungen die Kurzskalen zu sozialwissenschaftlich relevanten Kriteriumsvariablen aus den Bereichen Arbeit, Gesundheit, Politik und soziale Beziehungen aufweisen.

## 3 Methode

### *Stichproben*

Bei *Stichprobe 1* handelt es sich um eine Quotenstichprobe ($N$ = 539), die nach den Merkmalen Geschlecht, Alter, Bildung und Bundesland geschichtet ist. Befragt wurden deutschsprachige Personen ab einem Alter von 18 Jahren. Die Erhebung erfolgte in zwei Wellen mit einem zeitlichen Abstand von 6 bis 10 Wochen. An Welle 2 nahmen $N$ = 338 Befragungspersonen der Welle 1 teil. Die Daten wurden im Rahmen eines Interviews (CAPI; Computer Assisted Personal Interview) oder durch die Vorgabe eines Papierfragebogens („Papier-und-Bleistift"/Selbstausfüller) erhoben. *Stichprobe 2* mit $N$ = 1.134 Befragungspersonen ist eine Zufallsstichprobe, die repräsentativ für die Wohnbevölkerung in Deutschland über einem Alter von 18 Jahren ist. Sie wurde mithilfe des ADM-Stichprobensystems F2F („Random Route") der Arbeitsgemeinschaft deutscher Marktforschungsinstitute gezogen. Die Daten dieser Interviews wurden größtenteils im CAPI-Modus erhoben; der letzte Teil der Erhebung im CASI-Modus (CASI: Computer Assisted Self Interview). Tabelle 1 fasst die Charakteristika der beiden Stichproben zusammen.

### *Vorgehen*

Ausgangspunkt für die Entwicklung der Kurzskalen USS-8 waren die jeweils 10 Items umfassenden Originalskalen, die von Schmitt und Kollegen (2005) sowie von Schmitt, Baumert, Gollwitzer und Maes (2010) vorlegt wurden. Für jede der vier Perspektiven der Ungerechtigkeitssensibilität wurden jeweils zwei Items ausge-

Tabelle 1 Charakteristika der Stichproben

| | | Stichprobe 1 | | Stichprobe 2 |
|---|---|---|---|---|
| | | Welle 1 | Welle 2 | |
| *Stichprobe* | | | | |
| Umfang [*N*] | | 539 | 338 | 1.134 |
| Art | | Quote | Quote | Zufall |
| Modus | | CAPI, P&B | CAPI, P&B | CAPI, CASI |
| *Zusammensetzung* | | | | |
| Geschlecht [% Frauen] | | 52,5% | 52,1% | 55,6% |
| Alter [*M*(*SD*)] | | 47.2 (15.2) | 46.7 (15.1) | 53.3 (18.4) |
| Bildung | ≤ 9 Jahre | 44,7% | 45,3% | 37,2% |
| | 10 Jahre | 30,2% | 27,9% | 37,0% |
| | ≥ 11 Jahre | 23,7% | 25,4% | 25,8% |

*Anmerkungen. CAPI = Computer Assisted Personal Interview; CASI = Computer Assisted Self Interview; P&B = Papier-und-Bleistift (Selbstausfüller).*

wählt. Dabei musste entschieden werden, welche zwei der drei durch die Items der Originalskalen operationalisierten Indikatoren (spezifische emotionale Reaktion, unspezifische emotionale Belastung, Penetranz von Gedanken an eine widerfahrene Ungerechtigkeit) aufgenommen und welcher Indikator ausgeschlossen werden sollte. Als Kriterium diente der Grad der Einfachstruktur der Faktorladungsmatrizen als Indikator faktorieller Validität auf der Grundlage der Daten von Schmitt und Kollegen (2005) sowie Schmitt et al. (2010). Am besten ließ sich die Einfachstruktur für alle vier Perspektiven jeweils mit einem Item der spezifischen emotionalen Reaktion und der unspezifischen emotionalen Belastung erreichen. Pro Facette der Ungerechtigkeitssensibilität wurden auf diese Weise jeweils zwei Items ausgewählt, so dass insgesamt acht Items zur Messung der Ungerechtigkeitssensibilität in die Datenerhebung und -analysen aufgenommen wurden.

In einer ersten Testbefragung erwies sich die in der Originalskala (vgl. Schmitt et al. 2010) gewählte Instruktion für Umfragen in bildungsheterogenen Stichproben als kognitiv zu anspruchsvoll. Folglich wurde die Instruktion der Originalskalen im Rahmen eines Expertenreviews leicht verändert. Dabei wurden die kognitiven Anforderungen der Instruktion so reduziert, dass die Skalen in allgemeinen Bevölkerungsumfragen einsetzbar sind. Einige der in der Testbefragung interviewten Personen äußerten Schwierigkeiten, insbesondere im Hinblick auf die Beantwortung der Items zur Tätersensibilität. Die Befragten gaben an, dass sie

selbst eine solche in den Items geschilderte Situation noch nicht erlebt hätten und das Item folglich nicht beantworten könnten. Um dieses Problem zu entschärfen, wurde in der Instruktion der folgende Satz ergänzt: „Sollten Sie eine Situation noch nicht selbst erlebt haben, antworten Sie bitte so, wie Sie Ihrer Erwartung nach reagieren würden". In der CAPI-Bedingung der vorliegenden Untersuchung wurden den Befragten neben der Antwortskala auch die Items der Skala schriftlich präsentiert, um die Befragten zusätzlich kognitiv zu entlasten. Die Antwortskala der Items war sechsstufig mit den Endpunkten „trifft überhaupt nicht zu" (1) und „trifft voll und ganz zu" (6).

### *Messinstrumente*

Zur Validierung der neuen Kurzskalen wurden in Stichprobe 1 und 2 weitere psychologische Merkmale sowie sozialwissenschaftliche Inhaltsvariablen erhoben. Hierzu wurden etablierte Messinstrumente genutzt, z.B. zur Erfassung von Lebenszufriedenheit (SWLS, Diener/Emmons/Larsen/Griffin 1985; nur in Stichprobe 1), Kontrollüberzeugungen (Jakoby/Jacob 1999; nur in Stichprobe 1), Optimismus und Pessimismus (LOT-R, Glaesmer/Hoyer/Klotsche/Herzberg 2008; nur in Stichprobe 1), allgemeiner Selbstwirksamkeitserwartung (Schwarzer/Jerusalem 1999; nur in Stichprobe 1), zwischenmenschlichem Vertrauen (SOEPtrust; Naef/Schupp 2009) sowie zur Erfassung der Hauptdimensionen der Persönlichkeit nach dem Fünf-Faktoren-Modell (BFI-10, Rammstedt/John 2007).

An für die sozialwissenschaftliche Forschung relevanten Maßen wurden unter anderem die *Effort-Reward-Imbalance* (Siegrist et al. 2004; Siegrist et al. 2008), die physische und die psychische Beeinträchtigung der Gesundheit (Andersen/Mühlbacher/Nübling 2007), die politische Partizipation (in Anlehnung an Fragen B13 bis B19 im ESS 2008) und das delinquente Verhalten (Fragen A1 bis A4 im ALLBUS 2000) erhoben.

Darüber hinaus wurden ad hoc verschiedene Ein-Item-Skalen zur Messung unterschiedlicher Aspekte von Zufriedenheit konstruiert (Arbeitszufriedenheit, Zufriedenheit mit der eigenen Partnerschaft, Zufriedenheit mit der eigenen Gesundheit). Die bei GESIS entwickelten Kurzskalen PEKS und KSE-G wurden zur Messung der Political Efficacy sowie der sozialen Erwünschtheit eingesetzt (Beierlein/Kemper/Kovaleva/Rammstedt 2012; Kemper/Beierlein/Bensch/Kovaleva/Rammstedt 2012). Im Falle von Multi-Item-Skalen (vgl. Rammstedt 2004) wurden Summenwerte für diese Validierungsmaße gebildet.

Soziodemographische und -ökonomische Merkmale wie Geschlecht, Alter, Bildung, Wohnort und Einkommen wurden nach den Empfehlungen von Destatis (2010) abgefragt. In Anlehnung an die International Standard Classification of

Education (ISCED; UNESCO 1997) wurden die Befragten aufgrund ihrer Angabe zu ihrem höchsten Schulabschluss jeweils einer von drei Bildungsgruppen (niedrig, mittel, hoch) zugeordnet.

### *Statistische Datenanalyse*

Auf der Grundlage der bevölkerungsrepräsentativen Stichprobe 2 wurden die deskriptiven Kennwerte (Mittelwert, Standardabweichung, Spannweite, Trennschärfe, Schiefe, Exzess) der Verteilungen der einzelnen Items ermittelt.

Die Messgenauigkeit der neu entwickelten USS-8 Kurzskalen wurde auf Grundlage der Ladungen und Fehlervarianzen aus dem Messmodell einer Konfirmatorischen Faktorenanalyse (CFA; Brown 2006; Kline 2011) geschätzt. Als Schätzer wurde der Koeffizient ω von McDonald (1999: 90) herangezogen. Der Koeffizient gibt das Ausmaß an, in dem eine latente Variable von den Items geteilte Varianz reflektiert (Krohne/Hock 2007). Die Interpretation der Höhe von McDonald ω erfolgt analog zu Cronbach α. Die Stabilität der Skalen wurde mittels der Messwiederholungsdaten aus Stichprobe 1 berechnet. Die faktorielle Struktur der Kurzskalen wurde ebenfalls mittels CFA getestet. Folgende Indizes wurden dabei als Gütekriterien für den Modellfit herangezogen (vgl. Beauducel/Wittmann 2005; Brown 2006; Browne/Cudeck 1993; Hu/Bentler 1999): $\chi^2$ (*df*, *p*), Comparative Fit Index (CFI; zufriedenstellender Fit > .95), Tucker-Lewis-Index (TLI; zufriedenstellender Fit > .95), Root Mean Square Error of Approximation (RMSEA; zufriedenstellender Fit < .08).

Um die Vergleichbarkeit der Messergebnisse zwischen verschiedenen soziodemographischen Gruppen zu überprüfen, wurden Messinvarianzprüfungen auf der Basis von „Multiple Group Confirmatory Factor Analysis" (MGCFA) vorgenommen (Brown 2006; Steinmetz 2010). Als Schätzmethode wurde das Maximum-Likelihood-Verfahren gewählt. Für jede soziodemographische Variable wurden jeweils drei Modelle mit unterschiedlichen Parameterrestriktionen überprüft (vgl. Byrne 2009; Byrne/Shavelson/Muthén 1989). Um zu beurteilen, ob die Faktorenstruktur, die Itemladungen sowie die Itemintercepts zwischen den Gruppen äquivalent sind, wurden die Veränderungen des CFI als Kriterium herangezogen (Vandenberg/Lance 2000). Nach den Richtlinien von Cheung und Rensvold (1999) sowie Chen (2007) gelten Veränderungen des CFI von ≤ .01 als Hinweis darauf, die Invarianzhypothese nicht zurückzuweisen. Dabei ist jedoch zu beachten, dass ungleiche Stichprobengrößen zwischen den Gruppen dazu führen können, dass die Invarianzhypothese aufrechterhalten wird, obwohl tatsächlich keine Messäquivalenz vorliegt (Chen 2007).

Die konfirmatorischen Faktorenanalysen sowie die MGCFA wurden mit dem Programm M*plus* 6.1 (Muthén/Muthén 1998-2010) durchgeführt. Eine ausführliche Beschreibung des Vorgehens bei der Messinvarianzprüfung ist dem Beitrag von Beierlein, Kemper, Kovaleva und Rammstedt (im gleichen Band) zu entnehmen. Mittelwertunterschiede zwischen soziodemographischen Gruppen wurden auf der Basis der Ergebnisse der MGCFA durchgeführt, wenn zuvor die vollständige skalare Invarianz des Messinstruments für diese Gruppen etabliert werden konnte (Steinmetz 2010: 92 f.). Für den latenten Mittelwertvergleich wurde jeweils der latente Mittelwert einer Referenzgruppe auf Null fixiert und der globale Modellfit eines MGCFA-Modells getestet, bei dem die latenten Mittelwerte aller weiteren Gruppen mit dem Mittelwert der Referenzgruppe gleichgesetzt wurden (Allum/Read/Sturgis 2010: 41; Brown 2006: 282 ff.). Geprüft wurde dann, ob sich die beiden latenten Gruppenmittelwerte statistisch voneinander unterscheiden. Falls mehr als zwei Gruppen miteinander verglichen wurden, wurden mehrere Vergleiche durchgeführt, in dem jede Gruppe einmal als Referenzgruppe definiert wurde.

## 4 Ergebnisse

*Item- und Skalenstatistiken*

In Tabelle 2 sind die Item- und Skalenstatistiken der USS-8 Items getrennt nach Geschlecht, Alter und Bildungsstand aufgeführt. Alle Items weisen durchgehend zufriedenstellende Trennschärfen von $r_{it} > .63$ auf. Um mögliche Abweichungen von der Normalverteilung zu überprüfen, wurden Schiefe und Kurtosis der einzelnen Itemverteilungen berechnet (vgl. Tabachnik/Fidell 2013: 79). Die Ergebnisse zeigen, dass keine substanziellen Verteilungsauffälligkeiten zu berichten sind. Als Indikator für die psychometrische Itemschwierigkeit wurde der Mittelwert herangezogen (Bühner 2011: 219). Die höchsten Mittelwerte zeigten sich bei den Items, welche das Ausmaß der Tätersensibilität erfassen. Die Zustimmung zu diesen Items fiel den Befragten demzufolge besonders leicht.

Tabelle 2 Item- und Skalenstatistiken der acht Items der vier Ungerechtigkeitssensibilitäts-Kurzskalen (USS-8) in Stichprobe 2 ($N$ = 1.134)

| Skala/Item | *M* | *SD* | Schiefe | Kurtosis | $r_{it}$ |
|---|---|---|---|---|---|
| *Opfersensibilität* | 2.95 | 1.58 | 0.44 | -0.89 | - |
| 1. Es ärgert mich, wenn es anderen unverdient besser geht als mir. | 2.87 | 1.75 | 0.50 | -1.12 | .70 |
| 2. Es macht mir zu schaffen, wenn ich mich für Dinge abrackern muss, die anderen in den Schoß fallen. | 3.03 | 1.68 | 0.37 | -1.09 | .70 |
| *Beobachtersensibilität* | 3.46 | 1.37 | -0.05 | -0.72 | - |
| 3. Ich bin empört, wenn es jemandem unverdient schlechter geht als anderen. | 3.65 | 1.57 | -0.15 | -1.01 | .64 |
| 4. Es macht mir zu schaffen, wenn sich jemand für Dinge abrackern muss, die anderen in den Schoß fallen. | 3.29 | 1.45 | 0.09 | -0.85 | .64 |
| *Nutznießersensibilität* | 2.28 | 1.25 | 0.84 | 0.02 | - |
| 5. Ich habe Schuldgefühle, wenn es mir unverdient besser geht als anderen. | 2.27 | 1.37 | 0.96 | 0.06 | .75 |
| 6. Es macht mir zu schaffen, wenn mir Dinge in den Schoß fallen, für die andere sich abrackern müssen. | 2.27 | 1.30 | 0.90 | 0.12 | .75 |
| *Tätersensibilität* | 4.07 | 1.65 | -0.48 | -0.98 | - |
| 7. Ich habe Schuldgefühle, wenn ich mich auf Kosten anderer bereichere. | 4.23 | 1.71 | -0.64 | -0.89 | .79 |
| 8. Es macht mir zu schaffen, wenn ich mir durch Tricks Dinge verschaffe, für die sich andere abrackern müssen. | 3.90 | 1.79 | -0.33 | -1.26 | .79 |

*Anmerkungen. M = Mittelwert, SD = Standardabweichung, $r_{it}$ = Trennschärfe. sechsstufiger Antwortmodus: 1 = trifft überhaupt nicht zu, 6 = trifft voll und ganz zu. Der Range der Itemscores betrug für alle Items 1 bis 6.*

## *Reliabilität*

Zur Reliabilitätsschätzung wurde in Stichprobe 2 ein vierfaktorielles Modell getestet, bei dem die unstandardisierten Ladungen der je zwei Items auf dem gemeinsamen Faktor gleichgesetzt wurden. Dieser Restriktion liegt die Annahme zugrunde, dass die beiden Items das latente Konstrukt (bis auf Variationen des Messfehlers) gleichermaßen gut messen (Bühner 2011: 150). Die Faktorkorrelationen wurden dagegen frei geschätzt. Auf diese Weise wurde eine Höhe von McDonald ω von .82 für die Opfersensibilität, von .78 für die Beobachtersensibilität, von .85 für die Nutznießersensibilität sowie von .88 für die Tätersensibilität ermittelt. Demnach liegen die Reliabilitäten der neuen Kurzskalen zwischen .78 und .88 und damit unter denjenigen, die für die 10-Items umfassenden Langskalen berichtet werden (Schmitt et al. 2010: 220). Für Gruppenuntersuchungen können die hier für die

Kurzskalen berichteten Reliabilitäten jedoch als zufriedenstellend bewertet werden (Schermelleh-Engel/Werner 2012: 136). Neben der Reliabilität wurde auch die Stabilität der USS-8-Skalenwerte über die Korrelation der Messwerte in den beiden Wellen von Stichprobe 1 berechnet. Danach ergaben sich folgende Stabilitätskennwerte auf Ebene der manifesten Variablen: $r_{tt}$ = .56 für die Opfersensibilität, $r_{tt}$ = .44 für die Beobachtersensibilität, $r_{tt}$ = .54 für die Nutznießersensibilität und $r_{tt}$ = .47 für die Tätersensibilität. Der Anteil systematischer Varianz liegt dementsprechend zwischen 19 und 31 Prozent.

### *Messinvarianzprüfungen*

Die im Folgenden berichteten Messinvarianzprüfungen basieren auf dem oben beschriebenen vierfaktoriellen Modell. Getestet wurde die Äquivalenz dieses Messmodells in Gruppen von Personen mit unterschiedlichem Bildungsstand, Alter, Geschlecht und Wohnort (Region). Die Ergebnisse der Messinvarianzprüfungen sind in Tabelle 3 aufgeführt.

Insgesamt wurden vier separate Messinvarianzprüfungen für jede der soziodemographischen Variablen Bildung, Alter, Geschlecht und Region vorgenommen. Dabei wurden jeweils die Modellfits von drei getesteten, unterschiedlich restringierten Modellen (konfigurale, metrische und skalare Invarianz) miteinander verglichen. Für alle soziodemographischen Variablen zeigten die Modellfits jeweils aller drei Modelle eine akzeptable Passung der empirischen Daten auf das theoretische Modell (CFI ≥ .97; RMSEA ≤ .08). Der Fit der Modelle für die Bildungs-, Alters- und Geschlechtsgruppen veränderte sich nicht, wenn restriktivere Modelle getestet wurden. Demzufolge kann mit Hilfe der vorliegenden Studien die Annahme der konfiguralen, der metrischen sowie der skalaren Invarianz der USS-8 für verschiedene Bildungs-, Alters- und Geschlechtsgruppen sowie für Personen mit Wohnort im Westen und im Osten von Deutschland untermauert werden. Dies bedeutet, dass sich über die soziodemographischen Gruppen hinweg für alle Befragten bei gleicher latenter Merkmalsausprägung gleiche beobachtete Werte ergeben. Messinvarianz der USS-8 ist somit im Hinblick auf die getesteten Merkmale und Gruppen gegeben.

Tabelle 3 Ergebnisse der Prüfung dreier Invarianzmodelle für die USS-8 in unterschiedlichen soziodemographischen Gruppen auf Basis der repräsentativen Bevölkerungsstichprobe (Stichprobe 2)

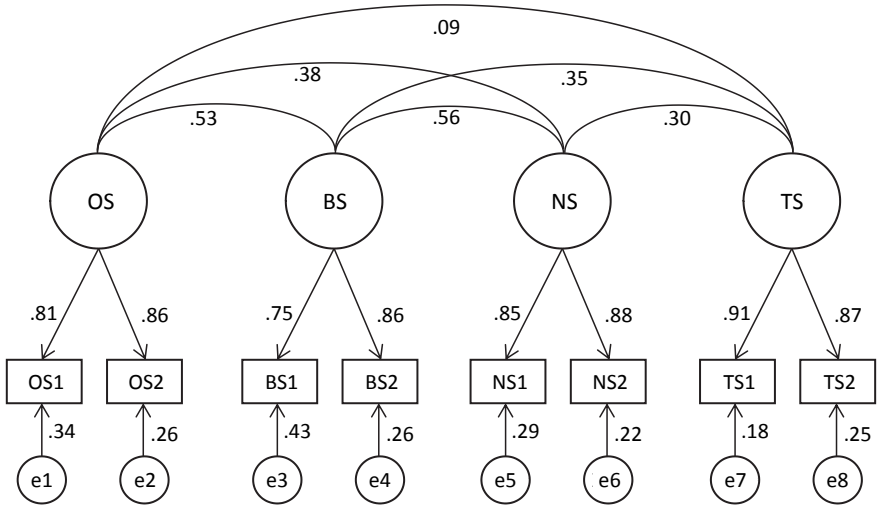
| | $\chi^2$ | *df* | *p* | CFI | TLI | RMSEA | Konfidenz-intervall | △ CFI |
|---|---|---|---|---|---|---|---|---|
| *Bildung* (niedrigster Bildungsstand $n_1$ = 422; mittlerer Bildungstand $n_2$ = 418; hoher Bildungstand $n_3$ = 290) | | | | | | | | |
| 1. Konfigurale Invarianz | 175.00 | 54 | .001 | .97 | .95 | .08 | (.07; .09) | - |
| 2. Metrische Invarianz | 186.29 | 62 | .001 | .97 | .96 | .07 | (.06; .09) | 0 |
| 3. Skalare Invarianz | 206.84 | 70 | .001 | .97 | .96 | .07 | (.06;.08) | 0 |
| *Alter* (18 bis 25 Jahre $n_1$ = 109; 26 bis 45 Jahre $n_2$ = 246; 46 bis 65 Jahre $n_3$ = 398; über 65 Jahre $n_4$ = 357) | | | | | | | | |
| 1. Konfigurale Invarianz | 192.82 | 76 | .001 | .97 | .96 | .07 | (.06; .09) | - |
| 2. Metrische Invarianz | 204.24 | 84 | .001 | .97 | .96 | .07 | (.06; .08) | 0 |
| 3. Skalare Invarianz | 218.10 | 96 | .001 | .97 | .96 | .07 | (.06; .08) | 0 |
| *Geschlecht* (männlich $n_1$ = 502; weiblich $n_2$ = 626) | | | | | | | | |
| 1. Konfigurale Invarianz | 126.38 | 36 | .001 | .98 | .96 | .07 | (.05; .08) | - |
| 2. Metrische Invarianz | 131.02 | 40 | .001 | .98 | .97 | .06 | (.05; .08) | 0 |
| 3. Skalare Invarianz | 139.31 | 44 | .001 | .98 | .97 | .06 | (.05; .07) | 0 |
| *Region* (West $n_1$ = 848; Ost $n_2$ = 280) | | | | | | | | |
| 1. Konfigurale Invarianz | 138.29 | 36 | .001 | .97 | .96 | .07 | (.06; .08) | - |
| 2. Metrische Invarianz | 142.93 | 40 | .001 | .97 | .96 | .07 | (.06; .08) | 0 |
| 3. Skalare Invarianz | 143.06 | 44 | .001 | .98 | .97 | .06 | (.05; .08) | .01 |

*Anmerkung. CFI = Comparative Fit Index, TLI = Tucker-Lewis-Index, RMSEA = Root Mean Square Error of Approximation.*

### *Faktorielle Validität*

Da Messinvarianz gegeben war, wurde die CFA auf Grundlage der gesamten Stichprobe 2 berechnet. Abbildung 1 zeigt das getestete konfirmatorische Faktorenmodell und gibt die empirischen Ergebnisse wieder. Das auf diese Weise spezifizierte Modell konnte die empirischen Daten angemessen beschreiben (Stichprobe 2: $\chi^2$ = 113.57, *df* = 18, *p* = .01, CFI = .98, TLI = .96, RMSEA = .07). Der globale Modellfit kann damit als zufriedenstellend bewertet werden. Die Höhe der standardisierten Faktorladungen beträgt danach stets mindestens $\lambda$ = .75, wonach ein großer Anteil der Varianz in der manifesten Variablen auf Unterschiede in der latenten Variable zurückführbar ist. Am niedrigsten fällt die Interkorrelation der Faktoren Täter- und Opfersensibilität aus (.09). Am stärksten sind die Faktoren Beobachtersensibilität

Abbildung 1 Faktorenmodell der Ungerechtigkeitssensibilitätskurzskalen (USS-8) in Stichprobe 2 (N = 1.134)




*In der Abbildung ist die standardisierte Lösung angegeben (OS = Opfersensibilität; BS = Beobachtersensibilität; NS = Nutznießersensibilität; TS = Tätersensibilität). e bezeichnet die Residuen.*

und Nutznießersensibilität miteinander korreliert (.56). Brown (2006: 32) zufolge weisen Faktorkorrelationen von .80 und höher auf eine mangelnde diskriminante Validität hin: In der vorliegenden Studie überschreitet keine der Faktorkorrelationen diese kritische Grenze.

### *Konstruktvalidität*

Bei der Konstruktvalidierung werden anhand theoretischer Überlegungen charakteristische Beziehungen der zu validierenden Skala mit anderen empirischen Indikatoren postuliert (Bühner 2011). Im Rahmen der Konstruktvalidierung der USS-8 wurde überprüft, ob aus der Fachliteratur bekannte typische psychologische Korrelate der Ungerechtigkeitssensibilität mit der USS-8 replizierbar sind. Tabelle 4 stellt die empirischen Ergebnisse der Konstruktvalidierung dar. Hierzu wurden multiple Regressionsanalysen durchgeführt und die standardisierten Regressionsgewichte der vier Ungerechtigkeitsperspektiven bei der Erklärung der jeweiligen Abhängigen Variablen berechnet. Das regressionsanalytische Vorgehen erlaubte es, den spezifischen Erklärungsbeitrag einer einzelnen Ungerechtigkeitsperspektive unter Kontrolle des Einflusses der jeweils verbleibenden Ungerechtigkeitsperspektiven darzustellen.

Tabelle 4 Validitätskoeffizienten der USS-8 als Ergebnisse multipler, linearer Regressionsanalysen (β-Koeffizienten, korrigierter Determinationskoeffizient $R^2_{korr}$) mit verschiedenen psychologischen Merkmalen als Abhängigen Variablen (AV) und den vier Perspektiven der Ungerechtigkeitssensibilität als Unabhängigen Variablen (S1 = Stichprobe 1; S2 = Stichprobe 2)

| | | Standardisierte Regressionsgewichte (β) | | | | |
|---|---|---|---|---|---|---|
| Psychologische Variable (AV) | Messinstrument | OS | BS | NS | TS | $R^2_{korr}$ |
| Big Five – Offenheit für Erfahrung | S2: Rammstedt/John (2007) | -.17** | .19** | .05 | .03 | .05 |
| Big Five – Gewissenhaftigkeit | S2: Rammstedt/John (2007) | -.13** | .06 | -.03 | .07* | .02 |
| Big Five – Extraversion | S2: Rammstedt/John (2007) | -.01 | .09* | -.17** | -.02 | .02 |
| Big Five – Verträglichkeit | S2: Rammstedt/John (2007) | -.16** | -.03 | .09 | .04 | .03 |
| Big Five - Neurotizismus | S2: Rammstedt/John (2007) | .05 | -.04 | .10** | .01 | .01 |
| Zwischenmenschliches Vertrauen | S1: Naef/Schupp (2009) | -.28** | .06 | -.08 | .12 | .07 |
| Lebenszufriedenheit | S1: Diener u.a. (1985) | -.37** | .20** | -.10 | .05 | .11 |
| Allgemeine Selbstwirksamkeit | S1: Schwarzer/Jerusalem (1999) | -.29** | .20** | -.17** | .04 | .08 |
| Optimismus | S1: Glaesmer u.a. (2008) | -.28** | .21** | -.08 | -.08 | .08 |
| Pessimismus | S1: Glaesmer u.a. (2008) | .27** | -.03 | .06 | -.07 | .06 |
| Kontrollüberzeugung - Internal | S1: Jakoby/Jacob (1999) | -.29** | .20** | -.14* | -.01 | .08 |
| Kontrollüberzeugung - External | S1: Jakoby/Jacob (1999) | .29** | -.06 | .04 | .04 | .08 |
| Soziale Erwünschtheit – Übertreibung positiver Qualitäten | S2: Kemper/Beierlein/Bensch/ Kovaleva/Rammstedt (2012) | -.21** | .02 | -.03 | .08* | .05 |
| Soziale Erwünschtheit – Untertreibung negativer Qualitäten | S2: Kemper/Beierlein/Bensch/ Kovaleva/ Rammstedt (2012) | -.34** | .13** | .01 | .14** | .12 |

*Anmerkungen. * p < .05, ** p <.01. OS = Opfersensibilität, BS = Beobachtersensibilität, NS = Nutznießersensibilität, TS = Tätersensibilität. Stichprobe 1 (Welle 1): CAPI und Selbstausfüller, N = 539; Stichprobe 2: CAPI, N = 1.134.*

Die Beziehungen der vier Perspektiven der Ungerechtigkeitssensibilität zu den *Big-Five*-Persönlichkeitsdimensionen werden im Folgenden mit den Befunden von Schmitt und Kollegen (2005; 2010) verglichen. Die Opfersensibilität stand in Übereinstimmung mit bisherigen Ergebnissen in einer negativen Beziehung mit der Verträglichkeit (β = -.16, $p < .01$). Über die Ergebnisse von Schmitt und Kollegen hinaus zeigten sich negative Zusammenhänge mit Offenheit für Erfahrung und Gewis-

senhaftigkeit. Im Gegensatz zu früheren Befunden erklärte die Opfersensibilität der Kurzskala keinen Anteil an Varianz in der Big-Five-Persönlichkeitsdimension Neurotizismus. Die Beobachtersensibilität stand erwartungsgemäß in einer positiven Beziehung mit den Big-Five-Dimensionen Offenheit für Erfahrung ($\beta = .19$, $p < .01$) und Extraversion ($\beta = .09$, $p < .01$). Entgegen früherer Befunde zeigten sich jedoch keine statistisch signifikanten Beziehungen zur Dimension Neurotizismus. Die Nutznießersensibilität stand dagegen in einem positiven Zusammenhang mit Neurotizimus ($\beta = .10$, $p < .01$) und in einem negativen mit Extraversion ($\beta = -.17$, $p < .05$). Zur Verträglichkeit zeigten sich keine positiven Beziehungen, obwohl dies theoretisch erwartet worden war. Die Tätersensibilität erklärte lediglich einen Anteil an der Varianz der Big-Five-Persönlichkeitsdimension Gewissenhaftigkeit ($\beta = .07$, $p < .05$). Dies war bereits in früheren Studien berichtet worden. Insgesamt konnten die vier Ungerechtigkeitsperspektiven zwischen 1% und 5% der Varianz in den Ausprägungen der Persönlichkeitsdimensionen aufklären ($.01 \leq R^2_{korr} \leq .05$). Die Abweichungen zu früheren Befunden werden im letzten Teil dieses Beitrags kritisch diskutiert.

Neben den Beziehungen zu den Big-Five-Persönlichkeitsdimensionen wurden auch Validitätskoeffizienten mit weiteren psychologischen Variablen berechnet. Erwartungsgemäß stand die Opfersensibilität in negativer Beziehung mit dem zwischenmenschlichen Vertrauen, der Lebenszufriedenheit, der allgemeinen Selbstwirksamkeitserwartung, dem Optimismus und der internalen Kontrollüberzeugung (siehe Tabelle 4). Die Beobachtersensibilität wies dagegen positive Beziehungen zur Lebenszufriedenheit, zur allgemeinen Selbstwirksamkeit, zum Optimismus und zur internalen Kontrollüberzeugung auf. Für die Nutznießersensibilität ergaben sich im Hinblick auf zwei Validitätskriterien ähnliche Beziehungen wie bei der Opfersensibilität: Diese Perspektive korrelierte negativ mit der Allgemeinen Selbstwirksamkeit und der internalen Kontrollüberzeugung. Die Effektstärken fielen jedoch schwächer aus als bei der Opfersensibilität. Die Beziehungen der Nutznießersensibilität zu zwischenmenschlichem Vertrauen und zu Pessimismus waren erwartungsgemäß nicht statistisch signifikant. Für die Tätersensibilität erbrachte die Studie keine statistisch signifikanten Ergebnisse hinsichtlich der Varianzaufklärung in den Kriterien. Insbesondere in Bezug auf Unterschiede in der Lebenszufriedenheit der Befragten konnten die Ausprägungen in den vier Perspektiven der Ungerechtigkeitssensibilität einen deutlichen Erklärungsbeitrag leisten ($R^2_{korr} = .11$; siehe Tabelle 4).

Im Hinblick auf die soziale Erwünschtheit untermauern die Ergebnisse die erwarteten theoretischen Beziehungen: Die Tätersensibilität zeigte positive Beziehungen mit der Antworttendenz der sozialen Erwünschtheit (insbesondere dem Aspekt der Untertreibung negativer Qualitäten: $\beta = .14$, $p < .01$). In den Items zur

Tätersensibilität wird der Befragte selbst als Verursacher von Ungerechtigkeit beschrieben. Dieses Verhalten gilt als sozial unerwünscht. Personen mit einer starken Neigung, negative Qualitäten der eigenen Person zu untertreiben, gaben daher eine höhere Tätersensibilität an. Die Opfersensibilität war dagegen substanziell negativ mit beiden Aspekten der sozialen Erwünschtheit assoziiert (Übertreibung positiver Qualitäten: $\beta = -.21$, $p < .01$; Untertreibung negativer Qualitäten: $\beta = -.34$, $p < .01$). Je höher die Neigung ist, einen guten Eindruck zu machen, desto niedriger war die Tendenz, hohe Kategorien bei der Opfersensibilität anzukreuzen.

Die Zusammenhänge der USS-8 mit sozialwissenschaftlichen Kriteriumsvariablen sind in Tabelle 5 ebenfalls als standardisierte Regressionskoeffizienten dargestellt. Um das Lesen zu erleichtern, wurden die einzelnen Kriteriumsvariablen den vier Bereichen Arbeit, Politik, Gesundheit, Soziale Beziehungen zugeordnet.

Im Bereich *Arbeit* war die Opfersensibilität erwartungsgemäß schwach, aber positiv mit der Effort-Reward-Imbalance ($\beta = .13$, $p < .05$) und negativ mit der Arbeitszufriedenheit ($\beta = -.17$, $p < .01$) assoziiert. Opfersensible nahmen ein stärkeres Ungleichgewicht zwischen den von ihnen erbrachten Leistungen auf der einen Seite und den Belohnungen für ihre Arbeit auf der anderen Seite wahr. Demgegenüber zeigten sich positive Beziehungen der Beobachter- und der Tätersensiblen mit der Arbeitszufriedenheit ($\beta = .12$ bzw. $\beta = .13$, $p < .05$). Auffällig ist der Befund, dass die Nutznießersensibilität negativ mit der Arbeitszufriedenheit zusammenhing ($\beta = -.15$, $p < .01$).

Auch in Bezug auf die Kriteriumsvariablen im Bereich *Politik* zeigten sich charakteristische Beziehungsmuster: Die Opfersensibilität stand in negativem Zusammenhang mit der internalen Komponente der Political Efficacy ($\beta = -.18$, $p < .01$). Hoch Opfersensible sprachen sich eine niedrige Kompetenz in der Beurteilung und aktiven Gestaltung politischer Prozesse zu. Dieses Ergebnis stimmt mit den negativen Beziehungen zur Allgemeinen Selbstwirksamkeit und zur internalen Kontrollüberzeugung überein (siehe Tabelle 4). Auch die Empfänglichkeit des politischen Systems für die Belange von Bürgerinnen und Bürgern wird von hoch opfersensiblen Personen als gering eingeschätzt ($\beta = -.10$, $p < .01$). Ein ähnliches Ergebnis zeigte sich auch für die Beobachtersensibilität (siehe Tabelle 5), jedoch mit dem bedeutenden Unterschied, dass eine hohe Beobachtersensibilität gleichzeitig mit niedriger Political Efficacy und hoher politischer Partizipation einhergehen kann: Je höher die Ungerechtigkeitssensibilität in dieser Komponente, desto stärker war die berichtete politische Teilhabe ($\beta = .13$, $p < .01$).

Die Beziehungen der vier Perspektiven der Ungerechtigkeitssensibilität mit Variablen aus dem Bereich *Gesundheit* entsprechen weitgehend den Erwartungen. Opfer- und Nutznießersensibilität zeigten in der Studie negative Zusammenhänge

Tabelle 5 Validitätskoeffizienten der USS-8 als Ergebnisse multipler, linearer Regressionsanalysen (β-Koeffizienten, korrigierter Determinationskoeffizient $R^2_{korr}$) mit verschiedenen sozialwissenschaftlichen Merkmalen als Abhängigen Variablen (AV) und den vier Perspektiven der Ungerechtigkeitssensibilität als Unabhängigen Variablen (Stichprobe 2; *N* = 1.134)

| | | Standardisierte Regressionsgewichte (β) | | | | |
|---|---|---|---|---|---|---|
| Sozialwissenschaftliche Variable (AV) | Messinstrument | OS | BS | NS | TS | $R^2_{korr}$ |
| *Arbeit* | | | | | | |
| Effort-Reward-Imbalance | Siegrist u.a. (2004) | .13* | -.04 | -.01 | .06 | .01 |
| Arbeitszufriedenheit | ad hoc konstruiert | -.17** | .12* | -.15** | .13* | .06 |
| *Politik* | | | | | | |
| Political Efficacy (Internal) | Beierlein/Kemper/Kovaleva/Rammstedt (2012) | -.18** | -.14** | -.01 | .02 | .03 |
| Political Efficacy (External)[1] | Beierlein/Kemper/Kovaleva/Rammstedt (2012) | -.10** | -.10** | .06 | .05 | .02 |
| Politische Partizipation | i.A. an ESS 2008 | -.07* | .13** | .06 | .07** | .03 |
| *Gesundheit* | | | | | | |
| Psychische Beeinträchtigung | Andersen u.a. (2007) | .08* | .06 | .13** | -.07 | .04 |
| Physische Beeinträchtigung | Andersen u.a. (2007) | -.05 | .03 | .07 | -.02 | .01 |
| Zufriedenheit mit der eigenen Gesundheit | ad hoc konstruiert | .04 | -.03 | -.10** | .11** | .01 |
| *Soziale Beziehungen* | | | | | | |
| Delinquentes Verhalten | ALLBUS 2000 | .23** | -.11** | .01 | -.07* | .05 |
| Zufriedenheit mit der eigenen Partnerschaft | ad hoc konstruiert | -.12** | .03 | -.07 | -.01 | .02 |

*Anmerkungen. * p < .05, ** p <.01. i.A. = in Anlehnung. OS = Opfersensibilität, BS = Beobachtersensibilität, NS = Nutznießersensibilität, TS = Tätersensibilität.[1]Ergebnis aus Stichprobe 1, Welle 1 (N = 539). Arbeitszufriedenheit in Anlehnung an SOEP 2009: „Wie zufrieden sind Sie gegenwärtig mit Ihrer Arbeit?"; Zufriedenheit mit der eigenen Gesundheit, ad hoc konstruiert: „Wie zufrieden sind Sie gegenwärtig mit Ihrer Gesundheit?"; Zufriedenheit mit Partnerschaft, ad hoc konstruiert: „Wie zufrieden sind Sie gegenwärtig mit Ihrer Beziehung zu Ihrem Partner/Ihrer Partnerin?". Für alle drei Zufriedenheitsskalen wurde eine 11-stufige Antwortskala mit 0 = „ganz und gar unzufrieden" bis 10 „ganz und gar zufrieden" gewählt. Nähere Informationen zu den Skalen können bei der Erstautorin angefordert werden (constanze.beierlein@gesis.org).*

mit der selbstberichteten psychischen Gesundheit (β = .08 bzw. β = .13, $p < .01$). Höhere Werte in diesen beiden Ungerechtigkeitssensibilitätsdimensionen gingen mit stärkeren selbstberichteten psychischen Beeinträchtigungen einher. Allerdings wirkte sich dies nur bei den Nutznießersensiblen auf die Zufriedenheit mit der eigenen Gesundheit aus (siehe Tabelle 5).

Auch für Kriteriumsvariablen im Bereich *soziale Beziehungen* sind theoriekonforme Befunde zu berichten. Die Opfersensibilität stand erwartungsgemäß in positiver Beziehung mit dem Ausmaß sozial abträglichen, delinquenten Verhaltens ($\beta = .23$, $p < .01$). Opfersensible berichteten z.B. häufigeres Schwarzfahren, Hinterziehen von Steuern sowie Autofahren mit mehr als 0.5‰ Alkohol im Blut. Hoch beobachtersensible und hoch tätersensible Personen zeigten dagegen eine geringere Neigung zu delinquentem Verhalten ($\beta = -.11$ bzw. $-.10$, $p < .05$). Darüber hinaus zeigte sich wie angenommen, dass Opfersensibilität und Zufriedenheit mit der eigenen Partnerschaft negativ assoziiert sind ($\beta = -.12$, $p < .01$). Hoch Opfersensible nahmen folglich ihre Partnerschaft als weniger zufriedenstellend wahr.

### *Soziodemographische Merkmale*

Im Rahmen der vorliegenden Studien wurde überprüft, ob sich die von Schmitt und Kollegen (2010) auf der Basis der Langskalen gefundenen Mittelwertunterschiede auch mit Hilfe der Kurzskalen der USS-8 abbilden lassen. Die beobachteten Skalenmittelwerte in den einzelnen soziodemographischen Gruppen sowie die Ergebnisse des Vergleichs der latenten Mittelwerte sind in Tabelle 6 dargestellt.

Hinsichtlich des *Geschlechts* zeigten sich statistisch signifikante Unterschiede in der Beobachter- sowie der Tätersensibilität: Beide waren bei Frauen höher ausgeprägt als bei Männern. Der Geschlechtseffekt bei der Tätersensibilität stimmt mit den Befunden zu den Langskalen überein. Allerdings fanden Schmitt und Kollegen (2010) zusätzliche Mittelwertunterschiede in der Nutznießer- sowie der Opfersensibilität.

Folgende Ergebnisse können für die soziodemographische Variable *Alter* berichtet werden: Nur im Hinblick auf die Opfersensibilität ist bei Unterschieden zwischen den Altersgruppen ein klarer Trend erkennbar: Wie bereits in der Studie mit den Langskalen gezeigt, nimmt die Opfersensibilität mit steigendem Alter ab. Schmitt und Kollegen (2010) berichten des Weiteren, dass Befragte unter 18 Jahren eine geringe Nutznießer- und die Tätersensibilität angaben als ältere Befragte. In der vorliegenden Studie konnten diese Unterschiede aufgrund des Mindestbefragungsalters von 18 Jahren nicht geprüft werden.

Die bisherigen Befunde zu *Bildungsunterschieden* wurden nur zum Teil repliziert. Erwartungsgemäß war die Beobachter-, die Nutznießer- und die Tätersensibilität in der Gruppe mit niedriger Bildung am geringsten ausgeprägt. Ein monotoner Trend über die drei Bildungsgruppen, wie von Schmitt und Kollegen (2010) berichtet, zeigte sich am deutlichsten bei der Tätersensibilität. Dabei muss beachtet werden, dass die hier verwendete Klassifikation des Bildungsgrads von der bei Schmitt und Kollegen (2010) genutzten abweicht.

Tabelle 6 Ergebnisse latenter Mittelwertvergleiche zwischen sozidemographischen Gruppen in den vier Ungerechtigkeitssensibilitätsperspektiven (in Anlehnung an Schmitt/Baumert/Gollwitzer/Maes 2010)

| | | Opfersensibilität | | | | Beobachtersensibilität | | | | Nutznießersensibilität | | | | Tätersensibilität | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | n | *M* | κ | S.E. | | *M* | κ | S.E. | | *M* | κ | S.E. | | *M* | κ | S.E. | |
| *Geschlecht* | | | | | | | | | | | | | | | | | |
| Männlich[1] | 502 | 2.99 | 0 | 0 | a) | 3.32 | 0 | 0 | a) | 2.23 | 0 | 0 | a) | 3.88 | 0 | 0 | a) |
| Weiblich | 626 | 2.94 | -0.07 | 0.07 | a) | 3.49 | 0.14 | 0.07 | b) | 2.35 | 0.10 | 0.07 | a) | 4.19 | 0.19 | 0.07 | b) |
| *Alter* | | | | | | | | | | | | | | | | | |
| 18-25 J.[1] | 109 | 3.80 | 0 | 0 | a) | 3.59 | 0 | 0 | a) | 2.48 | 0 | 0 | a) | 3.94 | 0 | 0 | a) b) |
| 26-45 J. | 264 | 3.16 | -0.34 | 0.13 | b) | 3.48 | -0.07 | 0.13 | a) | 2.36 | -0.06 | 0.12 | a) | 4.20 | 0.17 | 0.12 | a) |
| 46-65 J. | 398 | 3.00 | -0.44 | 0.12 | b) | 3.45 | -0.08 | 0.12 | a) | 2.31 | -0.12 | 0.12 | a) b) | 4.13 | 0.11 | 0.12 | a) b) |
| >65 J. | 357 | 2.46 | -0.89 | 0.12 | c) | 3.28 | -0.26 | 0.13 | b) | 2.17 | -0.26 | 0.12 | b) | 3.89 | -0.05 | 0.12 | b) |
| *Bildung* | | | | | | | | | | | | | | | | | |
| niedrig[1] | 420 | 2.85 | 0 | 0 | a) | 3.19 | 0 | 0 | a) | 2.13 | 0 | 0 | a) | 3.86 | 0 | 0 | a) |
| mittel | 418 | 3.09 | 0.17 | 0.08 | b) | 3.56 | 0.29 | 0.08 | b) | 2.39 | 0.26 | 0.08 | b) | 4.04 | 0.16 | 0.08 | b) |
| hoch | 290 | 2.84 | -0.08 | 0.08 | a) | 3.53 | 0.24 | 0.09 | b) | 2.40 | 0.23 | 0.08 | b) | 4.35 | 0.31 | 0.09 | c) |
| *Region* | | | | | | | | | | | | | | | | | |
| West[1] | 848 | 2.87 | 0 | 0 | a) | 3.35 | 0 | 0 | a) | 2.28 | 0 | 0 | a) | 4.09 | 0 | 0 | a) |
| Ost | 280 | 3.16 | 0.21 | 0.08 | b) | 3.60 | 0.20 | 0.08 | b) | 2.34 | 0.03 | 0.08 | a) | 3.94 | -0.10 | 0.08 | a) |

*Anmerkungen.* [1]*Referenzgruppe; der latente Mittelwert dieser Gruppe wurde auf Null fixiert. M = beobachteter Skalenmittelwert der Gruppe. Der theoretische Range der beobachteten Skalenmittelwerte betrug für alle vier Kurzskalen 1 bis 6.* κ *= geschätzter latenter Gruppenmittelwert; S.E. = geschätzter Standardfehler. Unterschiedliche Buchstaben in den Spalten zeigen statistisch signifikante Gruppenunterschiede an (*$p < .05$*).*

Statistisch signifikante Mittelwertunterschiede zwischen Bürgerinnen und Bürgern in *Ost- und Westdeutschland* konnten ausschließlich im Hinblick auf die Opfer- und die Beobachtersensibilität gefunden werden. In Übereinstimmung mit den Erwartungen berichteten Ostdeutsche im Durchschnitt stärkere Ausprägungen in diesen beiden Komponenten.

Das *persönliche Einkommen* einer Befragungsperson korrelierte statistisch signifikant negativ mit der Höhe der Skala Opfersensibilität der USS-8 ($r$ = -.14, $p$ = .001). Höhere Einkommen gingen mit einer niedrigeren Opfersensibilität einher; niedrigere Einkommen mit einer höheren Ausprägung dieses Persönlichkeitsmerkmals. Für die anderen drei Komponenten zeigten sich keine statistisch signifikanten Ergebnisse.

## 6 Diskussion

Ziel des vorliegenden Beitrags war es, vier neue Kurzskalen (USS-8) zur Erfassung der Sensibilität für Ungerechtigkeit im Hinblick auf ihren Nutzen für sozialwissenschaftliche Umfragen zu untersuchen (siehe auch Baumert und Kollegen, in Druck). Trotz der geringen Itemanzahl von jeweils zwei Items pro Skala erwiesen sich die Kurzskalen als hinreichend reliabel und valide. Die Stabilität bzw. Änderungssensitivität der neuen Kurzskalen wurde von Baumert und Kollegen (in Druck) auch mittels Latent-State-Trait-Analysen von Messwiederholungsdaten geschätzt. Dabei zeigte sich trotz der Kürze eine zufriedenstellende, geschätzte Reliabilität für alle Skalen von ≥ .80. Die geschätzte Stabilität (Trait Konsistenz) geben Baumert und Kollegen mit ≥ .43 an, die geschätzte Messgelegenheitsspezifität mit ≥ .33. Auch auf latenter Ebene zeigte sich, dass die USS-8 änderungssensitiver sind als die Langskalen (Schmitt et al. 2005).

Die Ergebnisse der Messinvarianzprüfung der USS-8 lassen darauf schließen, dass die Kurzskalen das Konstrukt der Ungerechtigkeitsensibilität in verschiedenen soziodemographischen Gruppen in vergleichbarer Qualität messen: Für Personen mit unterschiedlichem Bildungshintergrund, Alter, Geschlecht und Wohnort gilt, dass die Messung mit den neuen Kurzskalen bei gleicher latenter Merkmalsausprägung trotz des Unterschieds in diesen soziodemographischen Merkmalen zu gleichen Skalenwerten führt. Dies verdeutlicht, dass sich die USS-8 für den Einsatz in heterogenen Bevölkerungsumfragen in besonderer Weise eignen.

Die vorliegenden Studien liefern weitere Hinweise dafür, dass sich die vier Perspektiven nicht nur theoretisch zufriedenstellend voneinander trennen lassen, sondern auch empirisch (vgl. Brown 2006). Die vier Skalen der Ungerechtigkeits-

sensibilität weisen untereinander ein charakteristisches Beziehungsmuster auf, welches mit den Befunden von Schmitt und Kollegen (2010) übereinstimmt. Die Faktoren für die Täter- und die Opfersensibilität korrelierten erwartungsgemäß am niedrigsten miteinander. Diese beiden Perspektiven erfassen unterschiedliche Aspekte des Konstrukts, indem sich der Befragte einmal selbst in die Rolle des Opfers hineinversetzt und einmal in die Rolle des Täters. Die größte Gemeinsamkeit weisen die Beobachter- und die Nutznießersensibilität miteinander auf. Beobachter/-innen und Nutznießer/-innen können für Gerechtigkeit sorgen, indem sie auf eigene Vorteile verzichten und die Gerechtigkeit wiederherstellen (Schmitt et al. 2009).

Die Validitätskoeffizienten der USS-8 spiegeln zum Großteil die erwarteten Beziehungen zu weiteren psychologischen Variablen und zu sozialwissenschaftlichen Kriteriumsvariablen wider. Den Ergebnissen der Korrelationsanalysen der Ungerechtigkeitssensibilität mit den Big Five zufolge sind hoch Opfersensible wenig auf Reziprozität in sozialen Beziehungen bedacht, geben selten nach, sind wenig stressresistent und eher emotional labil (vgl. Gerlach et al. 2012). Dies zeigt sich unter anderem in der negativen Beziehung zur Big-Five-Dimension Verträglichkeit. Beobachtersensible zeigten sich in der Studie erwartungsgemäß offen für neue Erfahrungen und nehmen aktiv am gesellschaftlichen Geschehen teil. Dies kann auch umfassen, dass sie das soziale, politische und wirtschaftliche System kritisch hinterfragen und sich politisch engagieren. Dass Nutznießerensibilität mit Neurotizismus Gemeinsamkeiten aufweist, stimmt mit bisherigen Befunden ebenfalls überein. Neurotizismus kann neben Ängsten, Verärgerung und Traurigkeit auch Befangenheits- und Schamgefühle im Sinne eines existenziellen Schuldgefühls ausdrücken (vgl. Schmitt et al. 2009); letzteres empfinden insbesondere Nutznießersensible häufiger angesichts von eigenen Vorteilen. Hoch Tätersensible zeichnen sich zusätzlich durch eine hohe Gewissenhaftigkeit aus. Dies spiegelt wider, dass sie sich selbst an ihre ethischen Prinzipien und moralischen Maßstäbe halten.

Auch die Beziehungen zu weiteren psychologischen Merkmalen werfen ein charakteristisches Bild auf die mittels der Kurzskalen erfassten vier Perspektiven der Ungerechtigkeitssensibilität. Danach sind hoch Opfersensible eher misstrauisch, sind wenig zufrieden mit ihrem eigenen Leben, sind von ihrer Problemlösekompetenz nicht überzeugt und sehen ihr Leben nicht als in ihrer eigenen Kontrolle. Stattdessen sind sie stärker davon überzeugt, dass ihr Leben durch mächtige andere Personen oder den Zufall bestimmt wird. Im Gegensatz dazu schreiben hoch Beobachtersensible die Verantwortung für ihr eigenes Handeln nicht anderen Personen oder (un)glücklichen Umständen zu; sie sehen sich vielmehr selbst in der Verantwortung. Auch die Nutznießersensibilität lässt sich über den Zusam-

menhang mit psychologischen Variablen erneut von der Opfer- als auch von der Tätersensibilität abgrenzen. Hoch Nutznießersensible sind zwar genau wie Opfersensible ebenfalls weniger von ihren Handlungsmöglichkeiten überzeugt, ihr zwischenmenschliches Vertrauen ist jedoch unabhängig, d.h. nicht beeinträchtigt, von diesen ungünstigen Erwartungen.

Die Befunde zu den Kriteriumsvariablen aus den Bereichen Arbeit, Politik, Gesundheit und soziale Beziehungen verdeutlichen die Vorhersageleistung der Kurzskalen für sozialwissenschaftliche Kriterien. Hoch Opfersensible neigen auch im Arbeitsleben zu einem Misstrauen, ob sie tatsächlich das erhalten, was sie in ihren Augen verdienen. Sie erleben häufiger berufliche Gratifikationskrisen als weniger opfersensible Personen und dies führt möglicherweise langfristig zu Burnout und Depression (Hessler et al., in Vorbereitung). Opfersensible nahmen in den vorliegenden Studien eine höhere Diskrepanz zwischen beruflichen Anforderungen und Anerkennung am Arbeitsplatz (Effort-Reward-Imbalance) wahr. Gleichzeitig zeigte sich, dass hoch Opfersensible (wie auch die Nutznießersensiblen) eine geringere Arbeitszufriedenheit angaben. Dieses Ergebnis steht in Einklang mit Studien, nach denen die Effort-Reward-Imbalance mit einer geringeren Arbeitszufriedenheit assoziiert ist (vgl. Fietze, 2011; Siegrist et al. 2004).

Mit Hilfe der USS-8 konnten auch die erwarteten Zusammenhänge mit der psychischen Gesundheit repliziert werden. Personen mit hohen Ausprägungen in der Opfer- und Nutzersensibilität gaben häufiger an, sich niedergeschlagen zu fühlen und empfanden ihr Alltags- und Berufsleben stärker durch psychische Probleme beeinträchtigt. Dieses Ergebnis spiegelt den Zusammenhang dieser beiden Perspektiven der Ungerechtigkeitssensibilität mit Neurotizismus wider, den Schmitt und Kollegen (2010) in ihrer Publikation berichten.

Der negative Zusammenhang zwischen Opfersensibilität und Political Efficacy kann einerseits darauf zurückgeführt werden, dass die Aufmerksamkeit von Opfersensiblen stärker auf selbstbezogene Sorgen und Interessen gerichtet ist als auf gesamtgesellschaftliche Prozesse. Diese Annahme stimmt unter anderem mit dem von Fetchenhauer und Huang (2004) berichteten Verhaltenstendenzen von Opfersensiblen überein. Diese sind zwar stark daran interessiert, eigene Benachteiligung zu vermeiden. Gleichzeitig nehmen sie die Benachteiligung anderer und Normverletzungen in Kauf. Sich im Rahmen politischer Prozesse für Gerechtigkeit einzusetzen, liegt ihnen eher fern, da ihnen diese nicht unmittelbar zu Gute käme. Gleichzeitig drückt das Ergebnis zur Political Efficacy auch aus, dass Opfersensible sich selbst weniger in der Lage sehen, politische Prozesse zu verstehen und aktiv mit zu gestalten. Dies kann mit dem Eindruck der Ohnmacht einhergehen, „Opfer" einer Politik zu werden, über die sie selbst keine Kontrolle haben (vgl. Traut-Mattausch/

Guter/Zanna/Jonas/Frey 2011). Demgegenüber achten Beobachter- und Tätersensible auch darauf, dass anderen Personen Gerechtigkeit widerfährt. Eine Möglichkeit hierfür ist das aktive politische Engagement zugunsten von Gerechtigkeit (Beierlein/Werner/Preiser/Wermuth 2011).

Die USS-8 erwies sich auch als guter Prädiktor für Phänomene aus dem Bereich soziale Beziehungen. Da hoch Tätersensible in besonderem Maße geneigt sind, eine Bevorteilung der eigenen Person auf Kosten der Allgemeinheit zu vermeiden, zeigen sie weniger delinquente Verhaltensweisen wie Schwarzfahren oder Steuerhinterziehung. Das Gegenteil ist jedoch bei den Opfersensiblen der Fall. Die positive Korrelation mit der Delinquenz lässt darauf schließen, dass Opfersensible ihren eigenen Vorteil stärker im Blick haben als das Wohl der Allgemeinheit (vgl. Fetchenhauer/Huang 2004; Gollwitzer et al. 2005).

Die von Schmitt und Kollegen (2010) berichteten Unterschiede zwischen verschiedenen soziodemographischen Gruppen wurden in den vorliegenden Studien weitgehend repliziert. Es zeigten sich zum Beispiel erwartete Geschlechtsunterschiede in der Tätersensibilität, Altersunterschiede in der Opfersensibilität, sowie ein Einfluss der Bildung auf die Nutznießer- und die Tätersensibilität. Darüber hinaus erhärtete sich die Vermutung, dass der Wohnort (Ost- oder Westdeutschland) einen Einfluss auf die Opfer- und die Beobachtersensibilität hat. Zusammenfassend lassen sich auch mit den Kurzskalen zur Erfassung der Ungerechtigkeitssensibilität die erwarteten Unterschiede zwischen Bevölkerungsgruppen abbilden.

Folgende Einschränkungen der gegenwärtigen Studien sollten jedoch berücksichtigt werden und in zukünftigen Studien ausgeräumt werden:

1) Dass die Effektstärken der Validitätskoeffizienten mit den Big Five vergleichsweise niedrig ausfallen, kann u.a. darauf zurückgeführt werden, dass das in der Studie eingesetzte Big-Five-Inventar (BFI-10, Rammstedt/John 2007) nicht alle Facetten der Persönlichkeit gleichermaßen abbildet. Die Ergebnisse zu Zusammenhängen mit den Big Five sind folglich über Studien hinweg nur eingeschränkt vergleichbar. Insofern sollten in einer zukünftigen Studie vergleichbare Messinstrumente als Validierungskriterien verwendet werden.
2) Die Messinvarianz wurde ausschließlich über verschiedene soziodemographische Gruppen sichergestellt. Die Items und der Facettenansatz haben sich in Vorstudien als kognitiv anspruchsvoll erwiesen. Um die Anforderungen an die Befragten im Interview zu reduzieren, wurden ihnen neben der Antwortskala auch die Items auf Showcards präsentiert. Diese visuelle Unterstützung wäre in anderen Erhebungsmodi (z.B. Computer Assisted Telephone Interview) nicht möglich. Die Frage bleibt, ob die Messqualität an die spezifische Art der

Vorgabe gebunden ist und die Kurzskalen nur für bestimmte Erhebungsmodi empfohlen werden können.

3) Schließlich soll der Verzicht auf einen von drei Indikatoren der Ungerechtigkeitssensibilität kritisch diskutiert werden. Ziel war es ein Instrument vorzulegen, das mit zwei Items je Perspektive der Ungerechtigkeitssensibilität möglichst ökonomisch ist, ohne die Validität zu stark einzuschränken. Die Auswahl der Items für die Kurzskalen erfolgte unter Optimierung der Einfachstruktur der Items der vier Perspektiven anhand vorliegender Daten zur Originalskala. Die faktorielle Trennung der Perspektiven gelang mit jeweils einem Item zur Messung der Stärke spezifischer emotionaler Reaktionen und allgemein negativem Befinden als Konsequenz von Ungerechtigkeit. Inwiefern der Verzicht auf die Intrusivität von Gedanken an Ungerechtigkeit eine Einschränkung der Breite des erfassten Konstrukts und damit der Inhaltsvalidität des Messinstruments impliziert, muss empirisch überprüft werden. Dieser Befürchtung entgegengehalten können einerseits Befunde, dass die Indikatoren der Ungerechtigkeitssensibilität stark konvergieren (Schmitt et al. 1995). Andererseits weisen die vorliegenden Befunde darauf hin, dass die Validität des Messinstruments durch die Selektion der Indikatoren gegenüber den Originalskalen nur unwesentlich beeinträchtigt wurde. Denn im Hinblick auf Korrelationsmuster der Skalen untereinander sowie mit Validierungskonstrukten und –kriterien werden Ergebnisse der Originalskala weitgehend repliziert. Zukünftige Untersuchungen zur Vorhersage von kognitiven und verhaltensbezogenen Reaktionen auf Ungerechtigkeit sollten die inhaltliche Validität der Kurzskalen erneut überprüfen.

Zusammenfassend belegen die berichteten Ergebnisse die psychometrische Güte und den Mehrwert der neu entwickelten Kurzskalen USS-8 für die sozialwissenschaftliche Umfrageforschung.

## Literatur

Allum, N., S. Read u. P. Sturgis, 2010: Evaluating change in social and political trust in Europe using multiple group confirmatory factor analysis with structured means. S. 37-55 in: E. Davidov, P. Schmidt u. J. Billiet (Hg.): Cross-Cultural Analysis Methods and Applications. London, GB: Taylor & Francis.

Andersen, H. H., A. Mühlbacher u. M. Nübling, 2007: Die SOEP-Version des SF 12 als Instrument gesundheitsökonomischer Analysen. SOEP Papers on Multidisciplinary Panel Data Research, 6. Berlin: DIW.

Baumert, A., C. Beierlein, M. Schmitt, C. J. Kemper, A. Kovaleva, Liebig, S. u. B. Rammstedt, in Druck: Measuring four facets of Justice Sensitivity with two items each.

Beauducel, A. u. W. W. Wittmann, 2005: Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. Structural Equation Modeling 12: 41–75.

Beierlein, C., C. J. Kemper, A. Kovaleva u. B. Rammstedt, 2012: Ein Messinstrument zur Erfassung politischer Kompetenz- und Einflusserwartungen. Political Efficacy Kurzskala (PEKS).GESIS Working Paper, 18. Köln: GESIS.

Beierlein, C., C. S. Werner, S. Preiser u. S. Wermuth, 2011: Are just-world beliefs compatible with justifying inequality? Collective political efficacy as a moderator.Social Justice Research 24: 278-296.

Brown, T. A., 2006: Confirmatory factor analysis for applied research. New York: Guilford Press.

Browne, M. W. u. R. Cudeck, 1993: Alternative ways of assessing model fit. Sociological Methods & Research 21: 230–258.

Bortz, J. u. C. Schuster, 2010: Statistik für Sozial- und Humanwissenschaftler (7. Aufl.). Heidelberg: Springer.

Byrne, B., 2009: Structural equation modeling with AMOS: Basic concepts, applications, and programming (2nd edition). New York, NY: Taylor & Francis.

Byrne, B. M., R. J. Shavelson u. B. Muthén, 1989: Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin 105: 456–466.

Bühner, M., 2011: Einführung in die Test- und Fragebogenkonstruktion. PS Psychologie. München: Pearson Studium.

Chen, F. F., 2007: Sensitivity of goodness of fit indices to lack of measurement invariance. Structural Equation Modeling 14: 464-504.

Destatis, 2010: Statistik und Wissenschaft. Demographische Standards. Abgerufen am 01.11.2012 unter http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Publikationen/StatistikundWissenschaft/Band17__DemographischeStandards1030817109004,property=file.pdf

Diener, E., R. A. Emmons, R. J. Larsen u. S. Griffin, 1985: The satisfaction with life scale. Journal of Personality Assessment 49:7175.

Fetchenhauer, D., N. Goldschmidt, S. Hradil u. S. Liebig, 2010: Warum ist Gerechtigkeit wichtig? Antworten der empirischen Gerechtigkeitsforschung. München: Roman Herzog Institut.

Fetchenhauer, D. u. X. Huang, 2004: Justice Sensitivity and distributive decisions in experimental games. Personality and Individual Differences 36: 1015-1029.

Fietze, S., 2011: Arbeitszufriedenheit und Persönlichkeit: „Wer schaffen will, muss fröhlich sein!". SOEP Papers on Multidisciplinary Panel Data Research, 388. Berlin: DIW.

Gerlach, T. M., M. Allemand, D. Agroskin u. J. J. A. Denissen, 2012: Justice sensitivity and forgiveness in close interpersonal relationships: The mediating role of mistrustful, legitimizing, and pro-relationship cognitions. Journal of Personality 85: 1373-1413.

Glaesmer, H., J. Hoyer, J. Klotsche u. P. Y. Herzberg, 2008: Die deutsche Version des Life-Orientation-Tests (LOT-R) zum dispositionellen Optimismus und Pessimismus. Zeitschrift für Gesundheitspsychologie 16:2631.

Gollwitzer, M., T. Rothmund, A. Pfeiffer u. C. Ensenbach, 2009: Why and when Justice Sensitivity leads to pro- and antisocial behavior. Journal of Research in Personality 43: 999-1005.

Gollwitzer, M., M. Schmitt, R. Schalke, J. Maes u. A. Baer, 2005: Asymmetrical effects of Justice Sensitivity perspectives on prosocial and antisocial behavior. Social Justice Research 18: 183-201

Hessler, C., J. Pretsch, A. Hillert u. M. Schmitt (in Vorbereitung): Effects of justice sensitivity and effort-reward-imbalance on the mental health of teachers.

Hu, L. u. P. M. Bentler, 1999: Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling 6: 1–55.

Jakoby, N. u. R. Jacob, 1999: Messung von internen und externen Kontrollüberzeugungen. ZUMA-Nachrichten 45: 6171.

Kemper, C. J., C. Beierlein, D. Bensch, A. Kovaleva u. B. Rammstedt, 2012: Eine Kurzskala zur Erfassung des Gamma-Faktors sozial erwünschten Antwortverhaltens: Die Kurzskala Soziale Erwünschtheit-Gamma (KSE-G). GESIS Working Papers, 25. Köln: GESIS.

Kline, R. B., 2011: Principles and Practice of Structural Equation Modeling (3rd edition). London: Taylor & Francis.

Kraushaar, W., 2012: Der Aufruhr der Ausgebildeten. Vom Arabischen Frühling zur Occupy-Bewegung. Hamburg: Hamburger Edition.

Krohne, H. W. u. M. Hock, 2007: Psychologische Diagnostik: Grundlagen und Anwendungsfelder. Stuttgart: Kohlhammer.

Liebig, S., 2004: Empirische Gerechtigkeitsforschung: Überblick über aktuelle Modelle der psychologischen und soziologischen Gerechtigkeitsforschung. Humboldt-Universität Berlin: Arbeitsberichtsbericht Nr. 41 der Nachwuchsgruppe „Interdisziplinäre soziale Gerechtigkeitsforschung".

Liebig, S. u. H. Lengfeld (Hg.), 2002: Interdisziplinäre Gerechtigkeitsforschung. Zur Verknüpfung empirischer und normativer Perspektiven. Campus: Frankfurt a.M.

Liebig, S., H. Lengfeld u. S. Mau (Hg.), 2004: Verteilungsprobleme in modernen Gesellschaften. Campus: Frankfurt am Main.

McDonald, R. P., 1999: Test theory: A unified treatment. Mahwah: Erlbaum.

Montada, L., C. Dalbert, B. Reichle u. M. Schmitt, 1985: Urteile über Gerechtigkeit, „existenzielle Schuld" und Strategien der Schuldabwehr. S. 205 -225 in: F. Oser, W. Althoff u. D. Garz (Hg.): Moralische Zugänge zum Menschen – Zugänge zum moralischen Menschen. München: Kindt.

Muthén, L. K. u. B. Muthén, 1998-2010: Mplus User's Guide. Version 6. Los Angeles, CA: Muthén u. Muthén.

Naef, M. u. J. Schupp, 2009: Measuring trust: Experiments and surveys in contrast and combination. SOEPPapers on Multidisciplinary Panel Data Research No. 167. Berlin: DIW.

Noll, H.-H. u. S. Weick, 2012: Nicht einmal jeder Dritte empfindet soziale Unterschiede in Deutschland als gerecht. Informationsdienst Soziale Indikatoren (ISI) 48: 6-11.

Rammstedt, B., 2004: Zur Bestimmung der Güte von Multi-Item-Skalen: Eine Einführung. ZUMA How-to Berichte, Nr. 12.

Rammstedt, B. u. O. P. John, 2007: Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of Research in Personality 41:203212.

Schermelleh-Engel, K. u. C. Werner, 2012: Methoden der Reliabilitätsbestimmung. S. 120-141 in: H. Moosbrugger u. A. Kelava (Hg.): Testtheorie und Fragebogenkonstruktion. Heidelberg: Springer.

Schmitt, M., 1996: Individual differences in sensitivity to befallen injustice. Personality and Individual Differences 21: 3-20.

Schmitt, M., A. Baumert, D. Fetchenhauer, M. Gollwitzer, T. Rothmund u. T. Schlösser, 2009: Sensibilität für Ungerechtigkeit. PsychologischeRundschau 60: 8-22.

Schmitt, M., A. Baumert, M. Gollwitzer u. J. Maes, 2010: The Justice Sensitivity Inventory: Factorial validity, location in the personality facet space, demographic pattern, and normative data. Social Justice Research 23: 211-238.

Schmitt, M. u. M. Dörfel, 1999: Effects of justice sensitivity and procedural injustice in the workplace on job satisfaction and psychosomatic well-being. European Journal of Social Psychology 29: 443-453.

Schmitt, M., M. Gollwitzer, J. Maes u. D. Arbach, 2005: Justice sensitivity: Assessment and location in the personality space. European Journal of Psychological Assessment 21: 202-211.

Schmitt, M., R. Neumann u. L. Montada, 1995: Dispositional sensitivity to befallen injustice. Social Justice Research 8: 385-407.

Schmitt, M., J. Rebele, J. Bennecke u. N. Förster, 2008: Ungerechtigkeitssensibilität, Kündigungsgerechtigkeit und Verantwortlichkeitszuschreibungen als Korrelate von Einstellungen und Verhalten Gekündigter gegenüber ihrem früheren Arbeitgeber (Post CitizenshipBehavior). Wirtschaftspsychologie 10: 101-110.

Schwarzer, R. u. M. Jerusalem (Hg.), 1999: Skalen zur Erfassung von Lehrer- und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen. Berlin: Freie Universität Berlin.

Siegrist, J., D. Starke, T. Chandola, I. Godin, M. Marmot, I. Niedhammer u. R. Peter, 2004: The measurement of effort-reward imbalance at work: European comparisons. Social Science & Medicine 58:14831499.

Siegrist, J., N. Wege, F. Pühlhofer u. M. Wahrendorf, 2008: A short generic measure of work stress in the era of globalization: effort-reward imbalance. International Archives of Occupational and Environmental Health 82: 1005-1013.

Steinmetz, H., 2010: Estimation and Comparison of Latent Means Across Cultures. S. 87-118 in: E. Davidov, P. Schmidt u. J. Billiet (Hg.): Cross-Cultural Analysis Methods and Applications. London, GB: Taylor & Francis.

Tabachnick, B. G. u. L. S. Fidell, 2013: Using Multivariate Statistics (6th edition). Boston: Allyn and Bacon.

Traut-Mattausch, E., S. Guter, M. P. Zanna, E. Jonas u. D. Frey, 2011: When citizens fight back: Justice sensitivity and resistance to political reform. Social Justice Research24: 25-42.

United Nations Educational, Scientific and Cultural Organization (UNESCO), 1997: International Standard Classification of Education ISCED 1997. Paris: UNESCO.

Vandenberg, R. J. u. C. E. Lance, 2000: A Review and synthesis of the measurement invariance literature: Suggestions, practices and recommendations for organizational research. Organizational Research Methods 3: 4-69.

Wilkinson, R. u. K. Pickett, 2009: The Spirit Level. Why more equal societies almost always do better. London: Allan Lane.

| | |
|---|---|
| Anschrift der Autorin | Constanze Beierlein<br>GESIS – Leibniz-Institut für Sozialwissenschaften<br>B2,1<br>68159 Mannheim.<br>E-Mail: constanze.beierlein@gesis.org |
| Ko-Autoren/-innen | Anna Baumert<br>Fachbereich Psychologie<br>Universität Koblenz-Landau<br><br>Manfred Schmitt<br>Fachbereich Psychologie<br>Universität Koblenz-Landau<br><br>Christoph J. Kemper<br>Institut für Medizinische und<br>Pharmazeutische Prüfungsfragen (IMPP), Mainz<br><br>Beatrice Rammstedt<br>GESIS – Leibniz-Institut für Sozialwissenschaften<br>Mannheim |

# Appendix

## Ungerechtigkeitssensibilität-Kurzskalen (USS-8)

(Items 1 und 2 erfassen die Opfersensibilität, Items 3 und 4 die Beobachtersensibilität, Items 5 und 6 die Nutznießersensibilität, Items 7 und 8 die Tätersensibilität. Die numerischen Labels weichen geringfügig von den bei Schmitt und Kollegen [2010] verwendeten Labels ab.)

Menschen reagieren in unfairen Situationen sehr unterschiedlich. Im Folgenden möchten wir wissen, wie Sie selbst in unfairen Situationen reagieren. In den folgenden Aussagen werden verschiedene unfaire Situationen angesprochen. Bitte geben Sie an, wie sehr die jeweilige Aussage auf Sie zutrifft. Sollten Sie eine Situation noch nicht selbst erlebt haben, antworten Sie bitte so, wie Sie Ihrer Erwartung nach reagieren würden.

Zunächst geht es um Situationen, die zum Vorteil anderer und zu Ihrem Nachteil ausgehen.

| | trifft überhaupt nicht zu | | | | | trifft voll und ganz zu |
|---|---|---|---|---|---|---|
| (1) Es ärgert mich, wenn es anderen unverdient besser geht als mir. | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 | ☐ 6 |
| (2) Es macht mir zu schaffen, wenn ich mich für Dinge abrackern muss, die anderen in den Schoß fallen. | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 | ☐ 6 |

Nun geht es um Situationen, in denen Sie mitbekommen oder erfahren, dass jemand anderes unfair behandelt, benachteiligt oder ausgenutzt wird.

| | trifft überhaupt nicht zu | | | | | trifft voll und ganz zu |
|---|---|---|---|---|---|---|
| (3) Ich bin empört, wenn es jemandem unverdient schlechter geht als anderen. | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 | ☐ 6 |
| (4) Es macht mir zu schaffen, wenn sich jemand für Dinge abrackern muss, die anderen in den Schoß fallen. | ☐ 1 | ☐ 2 | ☐ 3 | ☐ 4 | ☐ 5 | ☐ 6 |

Hier geht es um Situationen, die zu Ihren Gunsten und zum Nachteil anderer ausgehen.

| | trifft überhaupt nicht zu | | | | | trifft voll und ganz zu |
|---|---|---|---|---|---|---|
| (5) Ich habe Schuldgefühle, wenn es mir unverdient besser geht als anderen. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 | □ 6 |
| (6) Es macht mir zu schaffen, wenn mir Dinge in den Schoß fallen, für die andere sich abrackern müssen. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 | □ 6 |

Zuletzt geht es um Situationen, in denen Sie selbst jemanden unfair behandeln, benachteiligen oder ausnutzen.

| | trifft überhaupt nicht zu | | | | | trifft voll und ganz zu |
|---|---|---|---|---|---|---|
| (7) Ich habe Schuldgefühle, wenn ich mich auf Kosten anderer bereichere. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 | □ 6 |
| (8) Es macht mir zu schaffen, wenn ich mir durch Tricks Dinge verschaffe, für die sich andere abrackern müssen. | □ 1 | □ 2 | □ 3 | □ 4 | □ 5 | □ 6 |

# Information for Authors

Methods, data, analyses (mda) publishes research on all questions important to quantitative methods, with a special emphasis on survey methodology. In spite of this focus we welcome contributions on other methodological aspects.

Manuscripts that have already been published elsewhere or are simultaneously submitted to other journals will not be considered. As a rule we do not restrict authors' rights. All rights remain with the author, and articles in mda are published under the CC-BY open-access license.

Mda aims for a quick peer-review process. All papers submitted to mda will first be screened by the editors for general suitability and then double-blindly reviewed by at least two reviewers. The decision on publication is made by the editors based on the reviews. The editorial team will contact the authors by email with the result at the latest eight weeks after submission; if the reviews have not been received by then, we provide a status update with a new target date.

When preparing a paper for submission, please consider the following guidelines:

- Please submit your manuscript by e-mail to mda(at)GESIS(dot)org.
- The total length of the manuscript shall not exceed 10.000 words.
- Manuscripts should...
  - be written in English, using American English spelling. Please use correct grammar and punctuation. Non-native English speakers should consider a professional language editing prior to publication.
  - be typed in a 12 pt Roman font, double-spaced throughout.
  - start with a cover page containing the title of the paper and contact details / affiliations of the authors, but be anonymized for review otherwise.
- Please also send us an abstract of your paper (approx. 300 words), a brief biographical note (no longer than 250 words), and a list of 5-7 keywords for your paper.
- Acceptable formats for Graphics are
  - Tiff
  - Jpeg (uncompressed, high quality)
  - EMF/WMF
- Please ensure a resolution of at least 300 dpi and take care to send hiqh-quality graphics. Line art images should have a resolution of 500-1000 dpi. Please note that we cannot print color images.
- The type area of our journal is 11.5 cm (width) x 18.5 cm (height). Please consider this when producing tables or graphics.
- Footnotes should be used sparingly.
- By submitting a paper to mda the authors agree to make data and program routines available for purposes of replication.

Please follow the APA guidelines when preparing in-text references and the list of references.

**Entire Book:**

Groves, R. M., & Couper, M. P. (1998). Nonresponse in household interview surveys. New York: John Wiley & Sons.

**Journal Article (with DOI):**

Klimoski, R., & Palmer, S. (1993). The ADA and the hiring process in organizations. Consulting Psychology Journal: Practice and Research, 45(2), 10-36. doi:10.1037/1061-4087.45.2.10

**Journal Article (without DOI):**

Abraham, K. G., Helms, S., & Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. American Journal of Sociology, 114(4), 1129-1165.

**Chapter in an Edited Book:**

Dixon, J., & Tucker, C. (2010). Survey nonresponse. In P. V. Marsden & J. D. Wright (Eds.), Handbook of Survey Research. Second Edition (pp. 593-630). Bingley: Emerald.

**Internet Source (without DOI):**

Lewis, O., & Redish, L. (2011). Native American tribes of Wisconsin. Retrieved April 19, 2012, from the Native Languages of the Americas website: www.native-languages.org/wisconsin.htm

For more information, please consult the Publication Manual of the American Psychological Association (Sixth ed.).