BERGISCHE
UNIVERSITÄT
WUPPERTAL

Dissertation im Fach Psychologie

mit dem Titel

Personality Understanding: Konzeptualisierung und Messung

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)

durch die Fakultät für Human- und
Sozialwissenschaften
der Bergischen Universität Wuppertal

vorgelegt von
Maike Pisters
aus Düsseldorf

angenommen von der Bergischen Universität Wuppertal
im August 2023

Danksagung

An dieser Stelle möchte ich mich bei allen Personen herzlich bedanken, die mich in den letzten Jahren dabei unterstützt haben, meine Dissertation anzufertigen. Mein ganz besonderer Dank gilt Prof. Dr. Ralf Schulze für die Möglichkeit, dieses Thema zu bearbeiten, die hilfreichen Rückmeldungen und seine umfangreiche Unterstützung in allen Phasen der Arbeit, aber insbesondere in den für mich schwierigsten und herausforderndsten Zeiten. Ebenso möchte ich mich bei Dr. Anna-Lena Jobmann und Dr. Susan Hellwig bedanken, die den Beginn beziehungsweise das Ende meiner Promotionszeit nicht nur durch das Teilen von wertvollen Erfahrungen, Ratschlägen und durch produktive Diskussionen begleitet und positiv geprägt haben.

Großer Dank gilt auch den vielen Personen, die mich im Rahmen ihrer Tätigkeiten als studentische oder wissenschaftliche Hilfskräfte sowie im Rahmen von Praktika oder Abschlussarbeiten bei den Vorbereitungen und Durchführungen der Studien unterstützt haben. Im besonderen Maße gilt dies für Markus Jansen, Julia Schaefer und Daniel Weppert. Vielen Dank auch an das Berufskolleg Elberfeld für die Möglichkeit, Datenerhebungen vor Ort durchführen zu können sowie an die zahlreichen Personen, die an den Studien teilgenommen haben.

Abschließend möchte ich mich bei meinen Eltern, meinem Bruder sowie bei Julia Flor, Kerstin Lindner und Aline Puzalowski für die emotionale Unterstützung und die vielen motivierenden Worte bedanken. Danke von ganzem Herzen, dass ihr immer an mich geglaubt habt!

Zusammenfassung

Beurteilungen der Persönlichkeit anderer Personen können richtig oder falsch sein, was unter anderem von der Fähigkeit der beurteilenden Person abhängt, akkurate Beurteilungen vorzunehmen. Auf Grund des Fehlens einer einheitlichen Konzeptualisierung dieser Fähigkeit wird ausgehend von vorhandenen Theorien und Modellen das Konstrukt Personality Understanding (PU) vorgeschlagen, das die Fähigkeit zum logischen Schlussfolgern über die Persönlichkeit anderer Personen beschreibt. Verfügbare Ansätze, die zur Erfassung von PU eingesetzt werden könnten, sind problembehaftet, da sie kein eindeutiges Scoring der Testpersonenantworten ermöglichen. Das Acquisition-Application (AcquA) Testdesign von Schulze und Roberts (2015) ist ein Ansatz zur Konstruktion von Leistungstests, der diese Scoring-Probleme lösen kann und sich zur Operationalisierung von PU eignet. In AcquA-Aufgaben zur Erfassung von PU wird zunächst präsentiert, mit welchen typischen Verhaltensweisen eine Zielperson auf bestimmte Ereignisse reagiert. In einer zweiten Phase wird die Zielperson in einer neuen, vergleichbaren Situation präsentiert und die Testpersonen müssen mögliche Reaktionen der Zielperson einschätzen. Hierdurch ist es möglich, die Testpersonenantworten eindeutig und logikbasiert als richtig oder falsch zu bewerten. Im Rahmen von drei empirischen Studien wurden AcquA-PU Aufgaben konstruiert, überarbeitet, die psychometrische Qualität untersucht und Validitätsevidenz gesammelt. In Studie 1 (N = 202) resultierte ein erwartungskonform großer positiver Zusammenhang der AcquA-PU Aufgaben mit AcquA-Aufgaben zur Erfassung von Emotional Understanding sowie maximal kleine Zusammenhänge mit Skalen der Big Five-Persönlichkeitsfaktoren. In Studie 2 (N = 204) konnte der erwartete große Zusammenhang mit einem Maß für logisches Schlussfolgern nicht aufgezeigt werden. Zudem ergaben sich vorläufige Hinweise auf einen möglichen Effekt der Ähnlichkeit zwischen Testperson und Zielperson bezüglich spezifischer Facetten der Verträglichkeit. In Studie 3 (N = 129 bzw. 162) wurde erneut der Zusammenhang mit logischem Schlussfolgern untersucht, der mittel ausfiel sowie höher als der Zusammenhang mit Merkfähigkeit. Außerdem zeigte sich ein nicht signifikanter Zusammenhang mit einem Maß für Social Understanding in unerwarteter Richtung. Auch der erwartete Zusammenhang mit akkuraten Persönlichkeitsbeurteilungen konnte nicht aufgezeigt werden. Es ergaben sich allerdings Hinweise auf mangelnde Validität der bestimmten Akkuratheitswerte. Insgesamt betrachtet deuten die Ergebnisse der drei Studien darauf hin, dass die konstruierten Aufgaben eine geeignete Grundlage darstellen, um zukünftig eine reliable und valide Erfassung von PU zu ermöglichen.

Abstract

Judgments about other people's personalities can be right or wrong, depending at least in part on the judging person's ability to make accurate judgments. Due to the lack of a consistent conceptualization of this ability, the construct Personality Understanding (PU) is proposed based on existing theories and models, which describes the ability to make correct inferences about other people's personalities. The main problem with available approaches that could be used to assess PU is that they do not enable an unequivocal scoring of the test takers responses. The Acquisition-Application (AcquA) Test Design by Schulze and Roberts (2015) is an approach for the construction of maximum performance tests that can address these scoring problems and is appropriate for the operationalization of PU. In AcquA tasks for the assessment of PU, the typical behavior with which a target person reacts to certain events is presented first. In a second phase, the target person is presented in a new, but similar situation and the test takers have to rate possible reactions of the target person. This allows for an unequivocal and logically derived scoring of the test takers responses. In three empirical studies, AcquA-PU tasks were constructed, revised, the psychometric quality was examined, and validity evidence was collected. In Study 1 (N = 202), consistent with expectations, the AcquA-PU tasks showed a large positive correlation with AcquA-tasks for the assessment of Emotional Understanding, as well as small correlations, at most, with scales of the Big Five personality traits. In Study 2 (N = 204), the assumed large correlation with a measure of reasoning ability could not be demonstrated. In addition, there was preliminary evidence of a possible effect of similarity between test taker and target person with respect to specific facets of agreeableness. In Study 3 (N = 129 or 162), the correlation with reasoning ability was examined again, which was medium and higher than the correlation to short-term memory. Furthermore, a nonsignificant relationship with a measure of Social Understanding was found in an unexpected direction. The assumed correlation with accurate personality judgments could not be demonstrated either. However, the results indicate validity issues of the calculated judgmental accuracy scores. Overall, the results of the three studies suggest that the constructed tasks provide a useful basis for a reliable and valid assessment of PU in the future.

Inhaltsverzeichnis

Elektronischer Anhang:

1. Einleitung

Beurteilungen der Persönlichkeit anderer Personen in Form von Einschätzungen und Vorhersagen über das typische Verhalten und Erleben finden im Alltag regelmäßig statt. Beispielsweise könnte man vor die Entscheidung gestellt werden, ob man eine Freundin um einen wichtigen Gefallen bittet oder nicht. Ob man sich letztendlich dafür oder dagegen entscheidet, hängt unter anderem davon ab, wie man die Gewissenhaftigkeit dieser Freundin beurteilt und ob man bei einer Zusage ihrerseits erwarten kann, dass sie einem auch wirklich helfen wird. Grundlage solcher Beurteilungen sind beispielsweise Überlegungen zu relevantem Verhalten aus der Vergangenheit: Hat die Freundin schon einmal einen Gefallen wie versprochen erledigt? Hat sie andere Versprechen eingehalten? Und auch bei weniger gut bekannten Personen, wie einem neuen Arbeitskollegen, kann es für ein erfolgreiches Zusammenarbeiten hilfreich sein, auf Basis von bereits beobachteten Reaktionen des Kollegen einzuschätzen, wie neugierig, fleißig, zurückhaltend, hilfsbereit oder reizbar er ist.

In der psychologischen Forschung sowie Praxis spielen Beurteilungen der Persönlichkeit anderer Personen in verschiedenen Bereichen ebenfalls eine wichtige Rolle. So liegt eines der bekanntesten Verfahren zur Erfassung bestimmter Persönlichkeitseigenschaften, die revidierte Fassung des NEO-Persönlichkeitsinventars nach Costa und McCrae (NEO-PI-R; deutschsprachige Version von Ostendorf & Angleitner, 2004), nicht nur in einer Selbstberichtsform (Form S), sondern zusätzlich auch in einer Fremdbeurteilungsform (Form F) vor. Für McCrae (1994) stellen diese beiden Formen sinnvolle gegenseitige Ergänzungen dar, die zusammen ein ausführlicheres Bild der Persönlichkeit liefern können. Laut Manual der deutschsprachigen Version des NEO-PI-R eignet sich die Form F in Situationen, in denen die zu beurteilende Person, die im Folgenden als *Zielperson* bezeichnet wird, zu keiner validen Selbstbeschreibung in der Lage ist (z.B. bei einer vorliegenden Behinderung oder bei bestimmten psychischen Störungen) oder wenn Antwortverfälschungen zu erwarten sind (Ostendorf & Angleitner, 2004). Vorgeschlagen wird im Manual auch der ergänzende Einsatz im Rahmen der Personalauswahl zur umfassenden Beschreibung der Persönlichkeit der Bewerber:innen. Darüber hinaus wurden im Rahmen der Konstruktvalidierung des NEO-PI-R Zusammenhänge zwischen beiden Formen erhoben und als konvergente und diskriminante Validitätsevidenz interpretiert (Ostendorf & Angleitner, 2004). Einsatzmöglichkeiten der Form F werden auch bei Piedmont (1998) thematisiert, der sich mit der Anwendung des NEO-PI-R im klinischen Kontext und der Forschung befasst hat. Der Autor geht in einem eigenen Kapitel auf die Anwendung der Fremdbeurteilungsform ein und beschreibt hier unter anderem die

Möglichkeit zur Validierung individueller Selbstberichte. Übereinstimmungen und Unterschiede zwischen Selbst- und Fremdbericht können seiner Ansicht nach wichtige Zusatzinformation im therapeutischen Kontext liefern. Darüber hinaus fokussiert Piedmont auf die Anwendung der Form F bei Ehepaaren zur Identifizierung und Evaluation von bestehenden Konflikten und thematisiert zudem den möglichen (zusätzlichen) Nutzen von Fremdberichten bei der Vorhersage von relevanten Außenkriterien.

Das NEO-PI-R ist nicht das einzige vielfach verwendete Instrument zur Erfassung von Persönlichkeitseigenschaften, das zusätzlich in einer Fremdbeurteilungsform vorliegt. Auch das HEXACO Personality Inventory Revised (HEXACO-PI-R; Ashton & Lee, 2009; Lee & Ashton, 2018) wurde sowohl in einer Selbstberichtsversion als auch in einer Fremdbeurteilungsversion konstruiert. Darüber hinaus gibt es Instrumente wie das California Q-Set (Block, 1961), die ursprünglich gezielt für die Beurteilung der Persönlichkeit anderer Personen entwickelt wurden. Die Nutzung einer Fremdbeurteilungsversion zur Sammlung von Validitätsevidenz im Rahmen der Testkonstruktion findet sich ebenfalls nicht ausschließlich beim NEO-PI-R, sondern wurde bereits vielfach in der Persönlichkeitsdiagnostik angewandt (z.B. Ashton & Lee, 2009; Goldberg, 1992; Lee & Ashton, 2018; Soto & John, 2017) und als besonders nützliches Validitätskriterium angesehen (McCrae, 1994).

Für die im Zusammenhang mit dem NEO-PI-R beschriebenen möglichen Anwendungsbereiche von Fremdbeurteilungen der Persönlichkeit sowie weitere Bereiche finden sich in der Literatur entsprechende Beispiele. So werden Fremdbeurteilungen der Persönlichkeit in der klinischen Psychologie als wichtige zusätzliche Informationsquelle bei der Erfassung von Persönlichkeitsstörungen diskutiert (Oltmanns & Turkheimer, 2009) und passende Fremdbeurteilungsverfahren wurden bereits entwickelt (Markon et al., 2013). Ein weiteres Beispiel ist bei Bagby et al. (1998) zu finden, die mit Hilfe der englischsprachigen Form F des NEO-PI-R die Validität der Selbstberichte von Patient:innen mit einer diagnostizierten Depression untersucht haben. Des Weiteren werden Fremdbeurteilungen als Kriterium verwendet, um zu erheben, ob und in welchem Ausmaß Personen ihre Persönlichkeitseigenschaften positiver einschätzen, als sie durch andere Personen wahrgenommen werden (Self-Enhancement; z.B. Asendorpf & Ostendorf, 1998; Kim et al., 2019). Im Bereich der Personalauswahl werden Fremdbeurteilungen bestimmter Persönlichkeitseigenschaften häufig im Rahmen von Einstellungsinterviews vorgenommen (Huffcutt et al., 2001; Levashina et al., 2014). Zusätzlich wurden bereits gezielte Trainings für die Interviewer:innen vorgeschlagen und deren Effekte untersucht (Powell & Bourdage, 2016; Powell & Goffin, 2009). Zur Erfassung der Persönlichkeit von Kindern und Jugendlichen wird

oftmals auf die Fremdbeurteilungen deren Eltern oder Lehrer:innen zurückgegriffen oder der Selbstbericht der Kinder und Jugendlichen hierdurch ergänzt (z.B. Brandt et al., 2021; Soto, 2016; Tackett, 2011). Zu diesem Zweck wurden ebenfalls entsprechende Instrumente entwickelt, wie beispielsweise eine Variante des bereits genannten California Q-Set (California-Child-Q-Set; dt. Version von Göttert & Asendorpf, 1989), das Inventory of Child Individual Differences (Deal et al., 2007; Halverson et al., 2003) oder das Hierarchical Personality Inventory for Children (deutsche Version von Bleidorn & Ostendorf, 2009). Darüber hinaus konnten Connelly und Ones (2010) in einer Metaanalyse prädiktive Validitätsevidenz von Fremdbeurteilungen bestimmter Persönlichkeitseigenschaften für die Außenkriterien akademische und berufliche Leistung aufzeigen sowie inkrementelle Validität über die selbstberichtete Persönlichkeit hinaus. Die Validitätskoeffizienten fielen hierbei zum Teil größer aus als die der selbstberichteten Persönlichkeit (Connelly & Ones, 2010).

Insgesamt zeigt sich also, dass Beurteilungen der Persönlichkeit anderer Personen Bestandteil verschiedenster Bereiche der psychologischen Forschung und Praxis sind. Ein wichtiger, aber nicht immer thematisierter Aspekt ist hierbei, dass die Zielperson eine bestimmte wahre Ausprägung auf der oder den interessierenden Persönlichkeitsdimension(en) aufweist beziehungsweise im Falle von Vorhersagen über das typische Verhalten und Erleben dieses zumindest theoretisch zeigen wird oder nicht. Hieraus folgt, dass die Persönlichkeitsbeurteilungen und Vorhersagen richtig oder falsch beziehungsweise unterschiedlich akkurat sein können (vgl. auch Back & Nestler, 2016; Funder, 1995). Unter Akkuratheit wird in der vorliegenden Arbeit in Anlehnung an Funder (1999; Funder & West, 1993) die Übereinstimmung zwischen Persönlichkeitsbeurteilung und wahrer Ausprägung der beurteilten Persönlichkeitseigenschaft(en) der Zielperson verstanden. Der Fokus liegt damit auf Beurteilungen individueller Zielpersonen.

Connelly und Ones (2010) konnten in einer Metaanalyse, in der die Akkuratheit unter anderem anhand der Übereinstimmung der Fremdbeurteilungen mit der selbstberichteten Persönlichkeit bestimmt wurde, aufzeigen, dass Fremdbeurteilungen der Persönlichkeit eine gewisse Akkuratheit besitzen. In der Literatur werden zudem verschiedene Aspekte diskutiert, oftmals unter Verweis auf das Realistic Accuracy Model von Funder (1995; vgl. Abschnitt 2.2.2), von denen die Akkuratheit der Fremdbeurteilungen abhängt. Hierzu zählen die Konsistenz im Verhalten der Zielperson, die Beobachtbarkeit der Persönlichkeitseigenschaft, die Quantität und Qualität der vorliegenden Information, aber auch eine Fähigkeit der Beurteiler:innen, eine akkurate Beurteilung vorzunehmen (Funder, 1995). Die praktische Relevanz einer solchen Fähigkeit ist im Hinblick auf die beispielhaft genannten Bereiche, in

denen Fremdbeurteilungen der Persönlichkeit eine wichtige Rolle spielen, leicht erkennbar. Das ist insbesondere dann der Fall, wenn individuelle Fremdberichte von einzelnen Beurteiler:innen verwendet werden, um beispielsweise zusätzliche Information für den psychotherapeutischen Prozess zu erheben oder um im Rahmen von Einstellungsinterviews zwischen mehreren Bewerber:innen zu entscheiden. Werden hierbei falsche Beurteilungen vorgenommen, kann das weitreichende individuelle Konsequenzen für die Patient:innen (z.B. unpassender Verlauf der Psychotherapie, Fehler bei Diagnostik einer Persönlichkeitsstörung) beziehungsweise Bewerber:innen (z.B. Absage, da fälschlicherweise davon ausgegangen wird, dass relevante Eigenschaft nicht vorliegt) haben. Es sollte in solchen Situationen daher das Ziel sein, die Beurteilungen von Personen vornehmen zu lassen, die das vergleichsweise gut können oder die individuelle Fähigkeitsausprägung der Beurteiler:innen zumindest in die Interpretation und Verwendung der Fremdbeurteilungen miteinzubeziehen. Die Relevanz einer solchen Fähigkeit steigt zudem in Situationen, in denen andere Aspekte, die sich ebenfalls auf die Akkuratheit auswirken könnten (z.B. Quantität der Information; Funder, 1995), nicht beeinflusst werden können.

Dass Fremdbeurteilungen der Persönlichkeit unterschiedlich akkurat sein können und dass dies unter anderem von einer Fähigkeit der Beurteiler:innen abhängt, wird in den dargestellten Anwendungsbereichen mehr oder weniger ausführlich thematisiert. Im deutschsprachigen Manual des NEO-PI-R wird beispielsweise erwähnt, dass die Übereinstimmung zwischen mehreren Beurteiler:innen von Aspekten wie der Beobachtbarkeit der Persönlichkeitseigenschaft abhängt (Ostendorf & Angleitner, 2004; vgl. auch Piedmont, 1998). Piedmont (1998) weist darauf hin, dass Fremdbeurteilungen fehlerhaft und verzerrt sein können und dass der Grund hierfür eine verzerrte Wahrnehmung der Beurteiler:innen sein kann. Eine den akkuraten Fremdbeurteilungen zugrunde liegende Fähigkeit der Beurteiler:innen wird aber weder im Manual des NEO-PI-R (Ostendorf & Angleitner, 2004) noch bei Piedmont (1998) diskutiert. Auch Kim et al. (2019) thematisieren im Zusammenhang mit der Kontrastierung von Fremd- mit Selbstberichten die Akkuratheit der Fremdbeurteilungen, wobei auch hier eine Fähigkeit unberücksichtigt bleibt. Ähnliches gilt für Oltmanns und Turkheimer (2009) im Zusammenhang mit der Diagnostik von Persönlichkeitsstörungen. Anders sieht es im Bereich Personalauswahl aus. Im Zusammenhang mit Einstellungsinterviews wird nicht nur die Akkuratheit von Fremdbeurteilungen der Persönlichkeit thematisiert und untersucht (z.B. Barrick et al., 2000; Nederström & Salmela-Aro, 2014; Schmid Mast et al., 2011), sondern auch Einflussfaktoren auf die Akkuratheit. Hierzu gehören einerseits Eigenschaften des Interviews, wie dessen Strukturiertheit

(Blackman, 2002), andererseits aber auch Eigenschaften der Interviewer:innen, wie eine Fähigkeit, akkurate Beurteilungen vornehmen zu können (Christiansen et al., 2005). Wie bereits erwähnt, wurden in diesem Zusammenhang auch bereits Trainings durchgeführt, um eine solche Fähigkeit und damit die Akkuratheit der Beurteilungen zu verbessern (Blanch-Hartigan & Hill Cummings, 2021; Powell & Bourdage, 2016; Powell & Goffin, 2009).

Insgesamt ist festzuhalten, dass interindividuelle Unterschiede in einer Fähigkeit, akkurate Beurteilungen der Persönlichkeit anderer Personen vornehmen zu können, Relevanz für die psychologische Forschung und Praxis besitzen. Damit diese Unterschiede auch berücksichtigt werden können, beispielsweise bei der Auswahl fähiger Beurteiler:innen, sind zwei Aspekte zwingend erforderlich: 1.) Die präzise Konzeptualisierung einer solchen Fähigkeit und 2.) eine daraus abgeleitete und dem Konstrukt angemessene Operationalisierung, die eine reliable und valide Erfassung dieser Fähigkeit erlaubt. Hinsichtlich beider Aspekte weist die bisherige Forschung allerdings noch Unklarheiten und Probleme auf.

Die Frage, welche Personen gute Beurteiler:innen sind, ist eine der ältesten im Bereich akkurater Persönlichkeitsbeurteilungen (Funder, 2012; für einen Überblick über die Geschichte der Forschung siehe Funder & West, 1993; Kenny, 1994). So haben sich bereits Adams (1927), Vernon (1933), Taft (1955) und Allport (1961) mit genau dieser Frage beschäftigt und zum Teil bereits eine Fähigkeit der Beurteiler:innen diskutiert. Mittlerweile finden sich in der Literatur mehrere konzeptuelle Modelle und Theorien, die sich mit dem Zustandekommen akkurater Beurteilungen und den Einflussfaktoren auf die Akkuratheit beschäftigen. Hierzu gehören das Linsenmodell (Brunswik, 1956), das darauf basierende Realistic Accuracy Model (Funder, 1995), das Weighted-Average Model (Kenny, 1991), das Truth and Bias Model (West & Kenny, 2011) sowie das Konzept der Dispositional Intelligence (Christiansen et al., 2005). Darüber hinaus liefern auch die Forschungsbereiche rund um die Soziale Intelligenz (z.B. Weis & Süß, 2005), die Kognitiv-Affektive Systemtheorie der Persönlichkeit (Mischel & Shoda, 1995) sowie verschiedene Attributionsmodelle (z.B. Gilbert et al., 1988; Trope, 1986) wertvolle Information für die Untersuchung akkurater Persönlichkeitsbeurteilungen. Von diesen Modellen und Theorien thematisieren allerdings nur wenige explizit eine Fähigkeit der Beurteiler:innen und das zum Teil mit unterschiedlichen Annahmen hinsichtlich der involvierten zentralen kognitiven Operation (z.B. Christiansen et al., 2005; Weis & Süß, 2005). Trotz der langen Historie existiert bis dato keine eindeutige und einheitliche Konzeptualisierung einer Fähigkeit, akkurate Persönlichkeitsbeurteilungen vornehmen zu können. Das erste Ziel der vorliegenden Arbeit bestand daher in der Konzeptualisierung und theoretischen Einordnung einer solchen Fähigkeit. Hierfür werden in Kapitel 2 konzeptuelle

Modelle und Theorien aus dem Forschungsbereich akkurater Persönlichkeitsbeurteilungen sowie relevanter angrenzender Forschungsbereiche vorgestellt und diskutiert. Die dort vorhandenen Beschreibungen einer entsprechenden Fähigkeit werden schließlich im ersten Abschnitt von Kapitel 3 zusammengefasst und soweit möglich integriert. Als Ergebnis wird das Konstrukt Personality Understanding vorgeschlagen, das als Fähigkeit zum logischen Schlussfolgern im Inhaltsbereich Persönlichkeit konzeptualisiert wird und von dem angenommen wird, dass es die zentrale kognitive Operation widerspiegelt, die akkuraten Beurteilungen der Persönlichkeit anderer Personen zugrunde liegt.

Weitere Heterogenität herrscht in der Literatur hinsichtlich der Bestimmung der Akkuratheit vom Persönlichkeitsbeurteilungen. Wie im Laufe von Kapitel 2 dargestellt wird, existieren verschiedene Ansätze, die zum Teil als austauschbare Varianten verwendet werden (vgl. Back & Nestler, 2016), aber konzeptuell und empirisch substantielle Unterschiede aufweisen (Hall et al., 2018). Zudem setzen sich bestimmte Akkuratheitswerte aus verschiedenen Komponenten zusammen, die unterschiedliche und unterschiedlich relevante Aspekte der Akkuratheit widerspiegeln (Biesanz, 2010; Cronbach, 1955). Darüber hinaus sind alle in der Literatur vorgeschlagenen Modelle und Methoden zur Bestimmung der Akkuratheit von einer sehr zentralen und grundlegenden Problematik betroffen, die Bernieri (2001) und Funder (1999) als Kriteriumsproblem bezeichneten. Versteht man unter Akkuratheit die Übereinstimmung zwischen Persönlichkeitsbeurteilung und wahrer Ausprägung der beurteilten Persönlichkeitseigenschaft(en) der Zielperson (Funder, 1999), so stellt sich die Frage nach der Bestimmung der wahren Ausprägung der eigentlich latenten Persönlichkeit. Hierfür wurden verschiedene Kriterien vorgeschlagen, die allerdings alle problembehaftet sind, da sie keine unverzerrte und damit keine eindeutige Bestimmung der wahren Ausprägung erlauben (Funder, 2012; Kenny, 1994). Diese beiden Problematiken erschweren die Operationalisierung des Konstruktes Personality Understanding, die ebenfalls ein Ziel der vorliegenden Arbeit darstellte und im zweiten Abschnitt von Kapitel 3 näher thematisiert wird. Wie noch näher erläutert wird, liegen sowohl im Hinblick auf die Konzeptualisierung als auch die Operationalisierung Ähnlichkeiten zwischen Personality Understanding und Emotional Understanding, einer Teilfähigkeit der Emotionalen Intelligenz (Mayer & Salovey, 1997; Mayer et al., 2016), vor. Unter anderem ist der Bereich der Emotionalen Intelligenz von einer analogen Kriteriumsproblematik betroffen, die in der Vergangenheit zu sehr ähnlichen Schwierigkeiten bei der Operationalisierung geführt hat (MacCann et al., 2004). Vor Kurzem wurde von Hellwig et al. (2020; siehe auch Schulze & Roberts, 2015) ein neuer Ansatz zur Erfassung von Emotional Understanding vorgeschlagen, der diese Problematik lösen soll und

in mehreren Studien zu vielversprechenden Ergebnissen geführt hat. Die Anwendung dieses Ansatzes zur Erfassung von Personality Understanding wurde im empirischen Teil der vorliegenden Arbeit vorgenommen. Hierfür wurden in insgesamt drei empirischen Studien Aufgaben zur Erfassung von Personality Understanding konstruiert, die psychometrische Qualität dieser Aufgaben untersucht sowie erste korrelative Validitätsevidenz gesammelt. Im Rahmen von Studie 1 (vgl. Kapitel 4) fand zunächst die Neukonstruktion von Aufgaben zur Erfassung von Personality Understanding in zwei verschiedenen Darstellungsformaten (2D-Grafiken vs. 3D-Simulationen) statt. Zudem wurde die psychometrische Qualität der neu konstruierten Aufgaben sowie die Zusammenhänge von Personality Understanding und Emotional Understanding sowie Personality Understanding und der Persönlichkeit der Testpersonen untersucht. Studie 2 (vgl. Kapitel 5) beinhaltete die Überarbeitung und Erweiterung des vorhandenen Pools der Aufgaben im 3D-Format und die Untersuchung des Zusammenhangs zwischen Personality Understanding und logischem Schlussfolgern. Ausgehend von den Ergebnissen der ersten Studie wurde zudem erneut der Zusammenhang zwischen Personality Understanding und der Persönlichkeit der Testpersonen betrachtet. In der dritten Studie (vgl. Kapitel 6) wurde schließlich ein erster Versuch unternommen, durch Überarbeitung von zwei Aufgaben die Aufgabenschwierigkeit gezielt zu steigern. Zudem wurde der Zusammenhang zwischen Personality Understanding und Social Understanding, der Akkuratheit von Persönlichkeitsbeurteilungen sowie Merkfähigkeit untersucht. Auf Basis der Ergebnisse der zweiten Studie wurde zudem erneut der Zusammenhang zwischen Personality Understanding und logischem Schlussfolgern betrachtet. In Kapitel 7 werden die Ergebnisse aller drei Studien zusammengefasst, bewertet, die Vor- und Nachteile der gewählten Operationalisierung von Personality Understanding diskutiert sowie ein Ausblick auf zukünftige Forschung gegeben. Kapitel 8 beinhaltet zudem eine allgemeine Zusammenfassung sowie ein allgemeines Fazit.

Die Ziele der vorliegenden Thesis bestanden kurz zusammengefasst somit in der Konzeptualisierung des Konstrukts Personality Understanding sowie in der Konstruktion und einer ersten Evaluation von Aufgaben zur Erfassung dieser Fähigkeit.

2. Akkurate Persönlichkeitsbeurteilungen

2.1 Persönlichkeit

Bevor akkurate Beurteilungen der Persönlichkeit anderer Personen sowie eine Fähigkeit der Beurteiler:innen näher betrachtet werden, soll zunächst der Inhalt solcher Beurteilungen thematisiert werden: Die Persönlichkeit und Persönlichkeitseigenschaften der Zielpersonen. Im Fokus stehen dabei die für die vorliegende Arbeit relevantesten Theorien und Modelle.

In der Literatur finden sich verschiedene Definitionen und Beschreibungen des Konstrukts Persönlichkeit, wie „personality is the dynamic organization within the individual of those psychophysical systems that determine his characteristic behavior and thought“ (Allport, 1961, S. 28), „personality is conceptualized as a stable system that mediates how the individual selects, construes, and processes social information and generates social behaviors“ (Mischel & Shoda, 1995, S. 246) oder „Persönlichkeit ist die nichtpathologische Individualität eines Menschen in körperlicher Erscheinung, Verhalten und Erleben im Vergleich zu einer Referenzpopulation von Menschen gleichen Alters und gleicher Kultur“ (Neyer & Asendorpf, 2018, S. 20). In einigen Ansätzen wird angenommen, dass sich Persönlichkeit aus verschiedenen Persönlichkeitseigenschaften zusammensetzt. Diese wurden definiert als „that which defines what a person will do when faced with a defined situation“ (Cattell, 1979, S. 14) oder als „dimensions of individual differences in tendencies to show consistent patterns of thoughts, feelings, and actions” (McCrae & Costa, 2003, S. 25). McCrae und Costa (2003) bezeichneten ihre Definition als phänotypisch, da sie beschreibt, wie Persönlichkeitseigenschaften aussehen und erkannt werden können. Trotz vorhandener Unterschiede darin, wie genau Persönlichkeit konzeptualisiert wird, stimmen die zitierten Definitionen und Beschreibungen in genau diesem Punkt überein. Allen ist zu entnehmen, dass sich die Persönlichkeit beziehungsweise Persönlichkeitseigenschaften einer Person unter anderem in ihren typischen Verhaltensweisen zeigen. Das heißt wiederum, dass aus den typischen Verhaltensweisen einer Person Rückschlüsse auf deren Persönlichkeitseigenschaften gezogen werden können (Costa & McCrae, 1992; Mischel & Shoda, 1995). Zur Untersuchung akkurater Persönlichkeitsbeurteilungen wurden in der Vergangenheit ganz unterschiedliche Verhaltensweisen und -indikatoren der Zielpersonen verwendet: Information aus direkten (Human & Biesanz, 2011) und indirekten Beobachtungen der Zielpersonen (Funder et al., 1995; Letzring, 2008), von den Zielpersonen geschriebene Texte (Borkenau et al., 2016),

Social Media Information (Darbyshire et al., 2016) und E-Mail-Adressen (Back et al., 2008) sowie Büro und Schlafzimmer der Zielpersonen (Gosling et al., 2002).

Im Laufe der Jahre wurden verschiedene Persönlichkeitsmodelle und -theorien vorgeschlagen, wobei laut John et al. (2008) in dem hierarchischen Fünf-Faktoren-Modell der Persönlichkeit (FFM) ein gewisser Konsens und eine allgemeine Taxonomie von Persönlichkeitseigenschaften gefunden wurde. Laut FFM kann die Persönlichkeit auf globaler, das heißt hierarchisch höchster Ebene durch fünf robuste und grundlegende Dimensionen beschrieben werden, die auch als *Big Five* bezeichnet wurden (Digman, 1990; Goldberg, 1993; McCrae & John, 1992; Tupes & Christal, 1992). Bei den fünf Dimensionen handelt es sich nach McCrae und John (1992) um *Offenheit für Erfahrungen* (Openness to Experience), *Gewissenhaftigkeit* (Conscientiousness), *Extraversion* (Extraversion), *Verträglichkeit* (Agreeableness) und *Neurotizismus* (Neuroticism). Auf hierarchisch niedrigerer Ebene finden sich je Big Five-Faktor zur differenzierteren Beschreibung der Persönlichkeit eine Reihe an spezifischeren Persönlichkeitseigenschaften, wobei verschiedene Anzahlen dieser sogenannten Facetten vorgeschlagen wurden (Costa & McCrae, 1992; Hofstee et al., 1992; Saucier & Ostendorf, 1999). Beispielsweise identifizierten Saucier und Ostendorf (1999) insgesamt 18 Facetten, Costa und McCrae (1992) postulierten je Big Five-Faktor sechs Facetten (vgl. auch Ostendorf & Angleitner, 2004) und Hofstee et al. (1992) schlugen neun Facetten je Faktor vor.

Trotz des relativ hohen Konsenses (siehe aber auch Goldberg, 1993) wurde das FFM hinsichtlich der Anzahl der globalen Persönlichkeitsdimensionen kritisiert und Modelle mit geringerer und höherer Anzahl vorgeschlagen. Eysenck (1992) war davon überzeugt, dass nur drei anstatt fünf globale Dimensionen unterschieden werden können – Psychotizismus, Extraversion, Neurotizismus – und Verträglichkeit, Gewissenhaftigkeit sowie Offenheit für Erfahrungen Faktoren einer hierarchisch niedrigeren Ebene darstellen. Seit einigen Jahren gibt es mit dem HEXACO-Modell (Ashton & Lee, 2007) zudem eine weitere Alternative zum FFM. In diesem Modell wird zusätzlich zu den Big Five die sechste globale Persönlichkeitsdimension *Ehrlichkeit-Bescheidenheit* (Honesty-Humility) postuliert. Allerdings handelt es sich bei dem HEXACO-Modell nicht ausschließlich um eine Erweiterung des FFM, da sich auch die Faktoren Neurotizismus und Verträglichkeit zwischen beiden Modellen inhaltlich unterscheiden (Ashton & Lee, 2007; Ashton et al., 2014).

In Ansätzen, die zur Beschreibung der Persönlichkeit Persönlichkeitseigenschaften heranziehen, wird oftmals angenommen, dass sich diese in Verhaltensweisen zeigen, die stabil über verschiedene Zeitpunkte sowie konsistent über verschiedene Situationen hinweg auftreten (McCrae & Costa, 2003; siehe auch Mischel & Shoda, 1995; Fleeson & Noftle, 2008). Daneben

gibt es Prozessansätze, die Variabilität im Verhalten über verschiedene Situationen hinweg explizit annehmen. Mischel und Shoda (1995) beschreiben in der Kognitiv-Affektiven Systemtheorie der Persönlichkeit diese als System, das sich aus mediierenden kognitiven und affektiven Einheiten (z.B. Erwartungen, Ziele, Werte, Emotionen) und deren Beziehungen und Organisation untereinander zusammensetzt. Dieses als stabil angenommene *Kognitiv-Affektive Persönlichkeitssystem* (Cognitive-Affective Personality System, CAPS; Mischel & Shoda, 1995) wird der Theorie zufolge durch Merkmale einer Situation aktiviert und generiert durch eine Interaktion der mediierenden Einheiten schließlich bestimmte Kognitionen, Affekte sowie Verhaltensweisen. Persönlichkeit zeigt sich diesem Ansatz nach auf behavioraler Ebene nicht nur in stabilen durchschnittlichen Ausprägungen von Verhaltensweisen über verschiedene Situationen hinweg, sondern zusätzlich auch in stabilen intraindividuellen Mustern von Beziehungen zwischen bestimmten Situationen und Verhaltensweisen (Mischel & Shoda, 1995). Diese wurden unter anderem auch als *Behavioral Signatures of Personality* oder *If…Then…Profiles* (Wenn-Dann-Profilen) bezeichnet (Kammrath et al., 2005; Mischel & Shoda, 1995; Shoda et al., 1994) und umfassen neben Verhalten auch die aktivierten Kognitionen und Affekte (Mischel & Shoda, 2008). Interessiert man sich beispielsweise für die Freundlichkeit einer Person, zeigt sich diese dem Ansatz zufolge nicht nur in der durchschnittlichen Freundlichkeit über verschiedene Situationen hinweg, sondern auch in einem spezifischen Muster, in welchen Situationen die Person freundlich reagiert und in welchen nicht. So kann es sein, dass zwei Personen dieselbe durchschnittliche Ausprägung von Freundlichkeit aufweisen, sich aber dennoch in ihrer Persönlichkeit unterschieden, da Person A nur in Situationen freundlich reagiert, in denen Person B es nicht tut und umgekehrt (vgl. Mischel & Shoda, 1995). Eine Invarianz über verschiedene Situationen hinweg nehmen Mischel und Shoda ebenfalls an, allerdings keine Invarianz des Verhaltens, sondern der gesamten Struktur des Persönlichkeitssystems. Ebenso lehnen sie das Konzept der Persönlichkeitseigenschaften nicht ab, sondern konzeptualisieren diese als stabile kognitiv-affektive Verarbeitungsstrukturen in Form von typischen Kognitionen, Affekten und Verhaltensstrategien – d.h. typischen kognitiv-affektiven Einheiten – sowie deren Organisation. Zudem werden die bereits beschriebenen Wenn-Dann-Profile als behaviorale Manifestation der durch relevante Aspekte einer Situation aktivierten Verarbeitungsstrukturen und damit der aktivierten Persönlichkeitseigenschaften angesehen (Mischel & Shoda, 1995). Die Wenn-Dann-Profile stellen somit Indikatoren der Persönlichkeitseigenschaften einer Person dar (Mischel & Shoda, 2008).

Zusammenfassend lässt sich sagen, dass unterschiedliche Konzeptualisierungen von Persönlichkeit und Persönlichkeitseigenschaften existieren, die sich unter anderem in der Rolle der Situation unterscheiden. Über eine lange Zeit herrschte in der Persönlichkeitspsychologie keine Einigkeit darüber, ob Verhalten vor allen durch die Persönlichkeitseigenschaften der Person oder vor allem durch die Situation bedingt wird (Fleeson & Noftle, 2008). Laut Fleeson und Noftle (2008) ist diese sogenannte Person-Situations-Debatte allerdings mittlerweile beendet. Die Autoren erläutern, dass die beiden konkurrierenden Ansichten zu einer gemeinsamen vereint werden können, indem Situationen in die Konzeptualisierung von Persönlichkeitseigenschaften mit aufgenommen und unterschiedliche Formen von Konsistenz betrachtet werden. Wie Fleeson und Noftle zudem herausstellen, ähnelt ihre Vorstellung zum Teil jener eben beschriebenen von Mischel und Shoda (1995). Einer der Unterschiede besteht allerdings darin, dass Fleeson und Noftle die Big Five als zentrale Persönlichkeitsdimensionen zur Beschreibung und Strukturierung der Persönlichkeitseigenschaften und des Verhaltens explizit anerkennen. Bei der Betrachtung einer Fähigkeit von Beurteiler:innen der Persönlichkeit anderer Personen kann die Rolle der Situation – nicht zuletzt auf Grund der Person-Situations-Debatte – nicht unbeachtet bleiben, was auch nicht der Fall ist (vgl. Christiansen et al., 2005; Letzring & Funder, 2021). Allerdings sollte die Konzeptualisierung und Operationalisierung einer solchen Fähigkeit möglichst unabhängig von einer konkreten theoretischen Vorstellung bezüglich der Rolle der Situation erfolgen, um keine Abhängigkeit von der Korrektheit dieser Ansicht zu erzeugen. Dieses Ziel wird im Rahmen der vorliegenden Arbeit verfolgt.

2.2 Konzeptuelle Modelle und Theorien

Es existieren eine Reihe an zum Teil sehr unterschiedlichen Theorien und Modellen, die sich mit akkuraten Persönlichkeitsbeurteilungen beschäftigen oder hierauf angewendet werden können. Diese unterscheiden sich in erster Linie dahingehend, ob sie Unterschiede in der Akkuratheit zwischen Beurteiler:innen und eine zugrunde liegende Fähigkeit thematisieren oder ob diese Aspekte nicht berücksichtigt werden. Letztere besitzen für die vorliegende Arbeit eine entsprechend geringere Relevanz. Zudem lassen sich die Theorien und Modelle grob in zwei Kategorien einteilen (vgl. Letzring & Funder, 2018): 1.) Konzeptuelle Modelle und Theorien, die Aussagen darüber treffen, wie akkurate Beurteilungen zustande kommen und von welchen Einflussfaktoren die Akkuratheit abhängt sowie 2.) Modelle und Methoden zur Bestimmung und Analyse der Akkuratheit von Persönlichkeitsbeurteilungen. Auch wenn diese

Aufteilung nicht trennscharf ist und es Modelle gibt, die in beide Kategorien eingeordnet werden können, wird sie soweit möglich im Folgenden übernommen. Die Modelle und Theorien der ersten Kategorie dienen hierbei als Grundlage hinsichtlich der Frage, wie eine Fähigkeit zu akkuraten Beurteilungen der Persönlichkeit anderer Personen konzeptualisiert werden kann. Modelle und Methoden der zweiten Kategorie stellen hingegen den Ausgangspunkt zur Beantwortung der Frage nach der angemessenen Operationalisierung einer solchen Fähigkeit dar und insbesondere, ob die aktuell verfügbaren Ansätze zur Bestimmung der Akkuratheit hierfür geeignet sind.

2.2.1 Linsenmodell

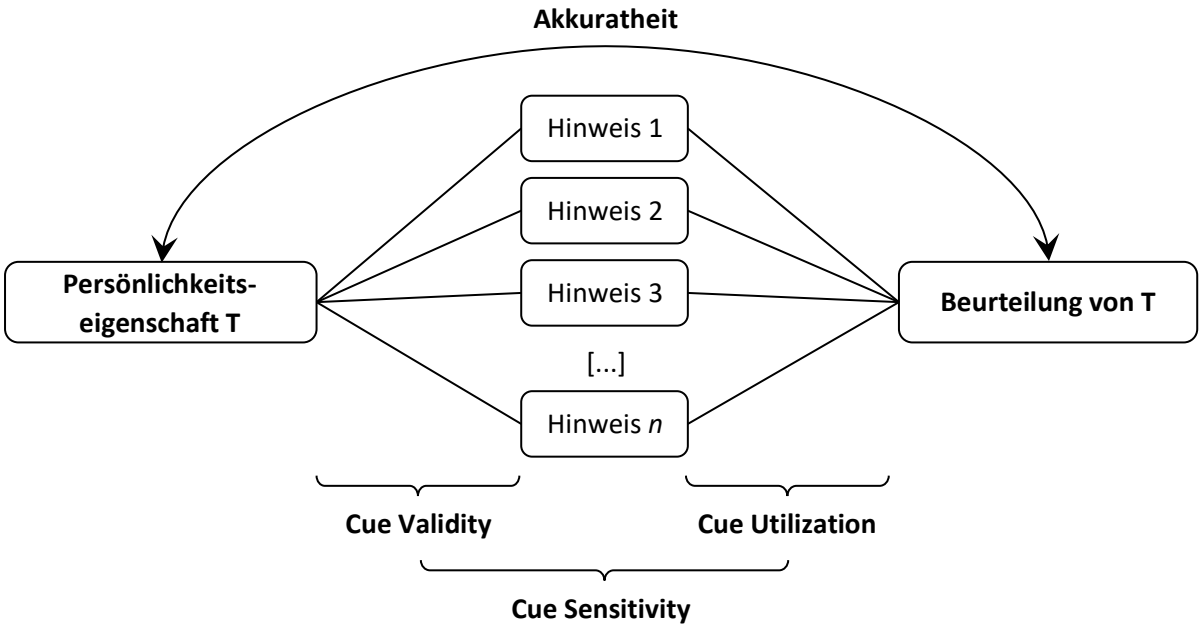
Ein Modell, das als allgemeines Rahmenkonzept verwendet werden kann, um die Akkuratheit von Persönlichkeitsbeurteilungen und die zugrundeliegenden Prozesse zu verstehen und zu erklären, ist das Linsenmodell von Brunswik (1956; vgl. Nestler & Back, 2013). Brunswik adressierte in seinem Modell allgemein die Wahrnehmung von distalen Variablen der Umgebung. Bei solchen Variablen kann es sich unter anderen auch um nicht direkt beobachtbare Eigenschaften oder Fähigkeiten von Personen, wie beispielsweise Intelligenz oder Persönlichkeit, handeln, die Brunswik (1956) als „covert distal“ (S. 6) bezeichnete. Die Wahrnehmung der distalen Variablen erfolgt dem Modell zufolge mit Hilfe von proximalen Variablen, die als für eine Person wahrnehmbare Hinweise auf die distalen Variablen beschrieben wurden (Brunswik, 1956). Die proximalen Hinweisvariablen sind dabei mehr oder weniger repräsentativ für die distale Variable (*Ecological Validity* oder *Cue Validity*; vgl. Letzring & Funder, 2018) und werden von der wahrnehmenden Person in unterschiedlichem Ausmaß für eine Reaktion auf die distale Variable, beispielsweise in Form von Einschätzungen der Intelligenz oder Persönlichkeit, verwendet (*Utilization*; Brunswik, 1956). Der Zusammenhang zwischen einer distalen Variablen und der Reaktion auf diese wurde von Brunswik als *Functional Validity* oder *(Perceptual) Achievement* bezeichnet.

Während Brunswik (1956) die Wahrnehmung der Persönlichkeit einer anderen Person bereits als Beispiel thematisierte, wurde die spezifische Anwendung und Erweiterung des Linsenmodells auf akkurate Persönlichkeitsbeurteilungen von Nestler und Back (2013) zusammenfassend beschrieben. Wenn beispielsweise die nicht direkt beobachtbare Gewissenhaftigkeit einer Zielperson beurteilt werden soll, nutzt die beurteilende Person in der Situation wahrnehmbare Hinweise auf die Gewissenhaftigkeit, um Rückschlüsse auf die Gewissenhaftigkeit zu ziehen und schließlich eine Beurteilung dieser vorzunehmen (Nestler & Back, 2013; vgl. Abbildung 2.1). Bei den wahrnehmbaren Hinweisen kann es sich unter

anderem um bestimmte Verhaltensweisen der Zielperson oder auch Aspekte ihrer Kleidung handeln (Nestler & Back, 2013). Auch in dieser spezifischen Anwendung des Linsenmodells unterscheiden Nestler und Back das Ausmaß, in dem ein Hinweis mit der Persönlichkeitseigenschaft zusammenhängt (Cue Validity) und das Ausmaß, in dem ein Hinweis für die Persönlichkeitsbeurteilung genutzt wird (Cue Utilization; vgl. Abbildung 2.1). Wie die Autoren zudem beschreiben, müssen für eine akkurate Persönlichkeitsbeurteilung valide Hinweise zur Verfügung stehen und diese dann auch von der beurteilenden Person genutzt werden. Darüber hinaus beschreibt die sogenannte *Cue Sensitivity* das Ausmaß, zu dem die beurteilende Person eher Gebrauch von validen anstatt invaliden Hinweisen macht (Nestler & Back, 2013).

Abbildung 2.1

Anwendung des Linsenmodells von Brunswik (1956) auf akkurate Beurteilungen der Persönlichkeit anderer Personen



Anmerkungen. Die Abbildung wurde in Anlehnung an die Darstellungen bei Back und Nestler (2016) sowie Nestler und Back (2013) erstellt. Die Akkuratheit entspricht der Functional Validity beziehungsweise (Perceptual) Achievement bei Brunswik (1956).

In einer Erweiterung des Linsenmodells zum dualen Linsenmodell (DLM) wird eine Dualität auf der Dimension Kontrolle versus Automatizität in den drei Elementen des Linsenmodells – das heißt bei der zu beurteilenden Persönlichkeitseigenschaft, den Hinweisen

auf die Persönlichkeitseigenschaft sowie der vorzunehmenden Persönlichkeitsbeurteilung – angenommen (Hirschmüller et al., 2013; Nestler & Back, 2013). Auf der Seite der Persönlichkeitseigenschaft wird im DLM zwischen explizitem und implizitem Selbstkonzept der Zielperson bezüglich ihrer eigenen Persönlichkeit unterschieden. Die Hinweise lassen sich dem Modell zufolge in eher kontrollierte Hinweise, die absichtlich und bewusst präsentiert werden (z.B. Kleidung), und automatischere Hinweise, die eher spontan gezeigt werden (z.B. behaviorale Reaktionen auf eine Situation), einteilen. Zudem wird auf Seite der Persönlichkeitsbeurteilung unterschieden zwischen bewussten Beurteilungen, an denen kognitive Prozesse höher Ordnung beteiligt sind, und intuitiven Beurteilungen, die kaum oder keine kognitiven Ressourcen in Anspruch nehmen (Hirschmüller et al., 2013). Diese duale Aufteilung der drei Elemente des Linsenmodells lässt die Untersuchung verschiedener Formen der Akkuratheit zu, je nachdem, ob der Fokus auf der bewussten oder intuitiven Beurteilung des expliziten oder impliziten Selbstkonzepts der Zielperson liegt (Hirschmüller et al., 2013; Nestler & Back, 2013).

Beim Linsenmodell handelt es sich nicht nur um ein konzeptuelles Modell, da es auch einen Rahmen zur Bestimmung und Analyse der Akkuratheit von Persönlichkeitsbeurteilungen liefert. Vorgeschlagen wurden auf Korrelations- und Regressionsanalysen basierende Koeffizienten zur Bestimmung der einzelnen Komponenten des Modells (u.a. Cue Validities, Cue Utilizations) inklusive der Akkuratheit (Stewart, 2001; Tucker, 1964) sowie ein Ansatz unter Verwendung von Strukturgleichungsmodellen (Nestler & Back, 2017).

Zum Linsenmodell in seiner ursprünglichen Version findet sich umfangreiche Forschung in verschiedenen Anwendungsbereichen (z.B. Hartwig & Bond, 2011; Karelaia & Hogarth, 2008; Kaufmann & Athanasou, 2009; Kaufmann et al., 2013; für eine Übersicht über verschiedene Forschungs- und Anwendungsbereiche siehe Hammond & Stewart, 2001). Dies macht die weite Verbreitung des Linsenmodells und dessen Rolle in der Forschung als allgemeines Rahmenkonzept für die Anwendung in unterschiedlichen Kontexten deutlich. Für den spezifischen und in der vorliegenden Arbeit fokussierten Bereich lassen sich ebenfalls eine Reihe an Anwendungsbeispielen finden. So war das Linsenmodell Grundlage für die Studie von Borkenau und Liebler (1992), in der die Akkuratheit von Persönlichkeitsbeurteilungen auf Basis physischer Merkmale untersucht wurde. Auch Naumann et al. (2009) nutzen das Linsenmodell als Rahmenkonzept zur Untersuchung von Persönlichkeitsbeurteilungen auf Basis des Aussehens der Zielpersonen. Weitere Beispiele sind die Studien von Back et al. (2008; Persönlichkeitsbeurteilungen auf Basis von E-Mail-Adressen), Küfner et al. (2010; Persönlichkeitsbeurteilungen auf Basis von durch die Zielperson verfassten kreativen Texten)

oder Tong et al. (2020; Persönlichkeitsbeurteilungen auf Basis von Selbstbeschreibungen in Online-Dating-Profilen). Allerdings ist das Linsenmodell auch nicht ganz frei von Kritik. So argumentieren Zebrowitz und Collins (1997), dass das Modell nur einen eingeschränkten Nutzen hat, da die Hinweise isoliert betrachtet und deren Konfiguration und Struktur außer Acht gelassen werden. Die Kritik bezieht sich dabei insbesondere auf die Anwendung des Linsenmodells zur Persönlichkeitsbeurteilung auf Basis physischer Merkmale wie Aussehen, Bewegungen und Stimme. Wiederum führen die Autor:innen auch an, dass, sobald theoretisch bedeutsame Konfigurationen ausfindig gemacht wurden, Brunswiks Linsenmodell eine geeignete Grundlage darstellt, Beurteilungen auf Basis dieser Konfigurationen näher zu untersuchen (Zebrowitz & Collins, 1997).

Das DLM wurde im Vergleich zu seiner ursprünglichen Variante bisher kaum eingesetzt. Auch wenn das Modell konzeptuell nachvollziehbar hergeleitet und die Annahme der Dualität in allen drei Elementen durch die Autor:innen begründet wird (vgl. Hirschmüller et al., 2013), fehlt es an empirischer Evidenz für das Modell im Ganzen. Diese ist derzeit begrenzt auf die zwei bei Hirschmüller et al. (2013) berichteten Studien. Eine weitere Einschränkung des DLM ist der Fokus auf das Selbstkonzept der Zielperson bezüglich ihrer eigenen Persönlichkeit. Zu beachten ist, dass das explizite Selbstkonzept nicht mit der wahren Ausprägung der Persönlichkeit übereinstimmen muss, da es bei der Operationalisierung über Selbstberichte (Hirschmüller et al., 2013) auf Grund von Selbst- und Fremdtäuschung zu verzerrten Angaben kommen kann (Asendorpf et al., 2002; Funder, 2012). Das implizite Selbstkonzept soll von solchen Verzerrungen zwar weniger betroffen sein (Asendorpf et al., 2002), allerdings wird beim DLM nicht thematisiert, inwieweit durch das explizite oder auch implizite Selbstkonzept die wahre Persönlichkeit abgebildet wird, was die Interpretation der Akkuratheit von Persönlichkeitsbeurteilungen, die auf Basis des Modells untersucht werden, einschränkt. Die Verwendung von Selbstberichten der Zielpersonen zur Operationalisierung der wahren Ausprägung der Persönlichkeitseigenschaft ist allerdings kein spezifisches Problem des DLM. Es handelt sich hierbei um ein grundlegendes und sehr zentrales Thema im Forschungsbereich akkurater Persönlichkeitsbeurteilungen (vgl. Abschnitt 3.2) und betrifft ebenso viele Anwendungen der ursprünglichen Variante des Linsenmodells (z.B. Back et al., 2008; Borkenau & Liebler, 1992).

Im Hinblick auf den Fokus der vorliegenden Arbeit besteht der größte Nachteil beider Modelle darin, dass Unterschiede zwischen Beurteiler:innen in der Akkuratheit von Persönlichkeitsbeurteilungen nur eingeschränkt erklärt werden können. Zum einen erlauben beide Modelle zwar prinzipiell die Erfassung interindividueller Unterschiede, allerdings

werden diese in Studien, die auf dem ursprünglichen Linsenmodell basieren, nicht immer betrachtet (z.B. Back et al., 2008; vgl. auch Hirschmüller et al., 2013). Zum anderen liefert das Linsenmodel zwar Anhaltspunkte zur Erklärung interindividueller Unterschiede, wie die unterschiedliche Nutzung von validen Hinweisen (Nestler & Back, 2013), allerdings werden weitere Einflussfaktoren, wie beispielsweise eine Fähigkeit, nicht thematisiert.

Die Betrachtung solcher Einflussfaktoren beziehungsweise möglicher Moderatoren der Akkuratheit finden sich dafür im Realistic Accuracy Model von Funder (1995), das unter anderem auf dem Linsenmodell von Brunswik basiert (Funder, 1995, 2012) und das daher oftmals als Variante des Linsenmodells bezeichnet wird (z.B. Back & Nestler, 2016).

2.2.2 Realistic Accuracy Model

Das Realistic Accuracy Model (RAM) von Funder (1995) ist ein Prozessmodell, das sich explizit mit dem Zustandekommen akkurater Persönlichkeitsbeurteilungen beschäftigt. Eines der Hauptziele des RAM besteht darin, mit Hilfe weniger Prozessvariablen zu erklären, wie akkurate Beurteilungen zustanden kommen und welche Faktoren sich auf die Akkuratheit auswirken (Funder, 1995). Unterschieden werden im RAM vier Prozessvariablen beziehungsweise Schritte, die stets für eine akkurate Persönlichkeitsbeurteilung erfolgreich absolviert werden müssen und im Folgenden auf Basis von Funder (1995, 1999) beschrieben werden (vgl. auch Abbildung 2.2):

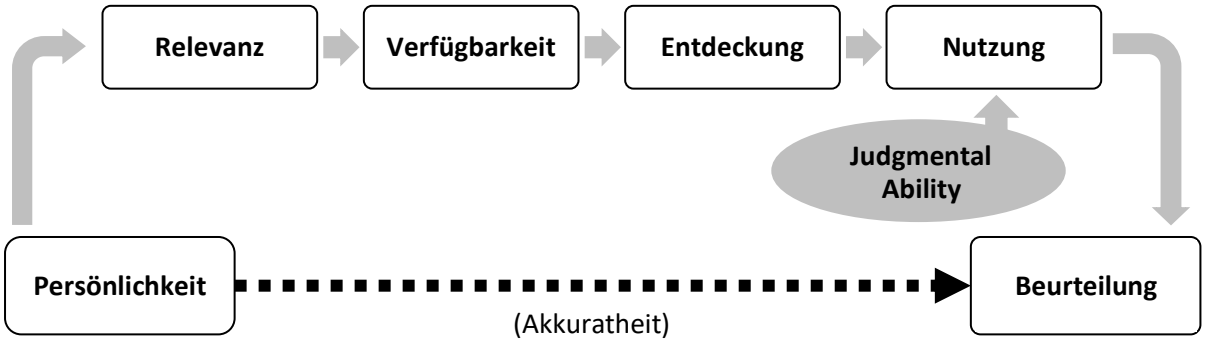
1. Relevanz: Die zu beurteilende Persönlichkeitseigenschaft muss sich im Verhalten zeigen. Das heißt, die Zielperson muss etwas Persönlichkeitsrelevantes tun.
2. Verfügbarkeit: Das persönlichkeitsrelevante Verhalten muss für die Beurteiler:innen verfügbar sein. Diesen muss es also prinzipiell möglich sein, das Verhalten wahrnehmen zu können.
3. Entdeckung: Das verfügbare persönlichkeitsrelevante Verhalten muss durch die Beurteiler:innen bewusst oder unbewusst wahrgenommen werden. Es darf nicht sehr schwer zu erkennen sein und die Beurteiler:innen dürfen beispielsweise nicht unaufmerksam sein.
4. Nutzung: Das verfügbare und entdeckte persönlichkeitsrelevante Verhalten muss durch die Beurteiler:innen korrekt genutzt und interpretiert werden. Es müssen von dem Verhalten Rückschlüsse auf die korrekte Persönlichkeitseigenschaft gezogen werden.

Während die ersten beiden Schritte (Relevanz und Verfügbarkeit) nur von der Zielperson abhängen, sind die beiden letzten Schritte (Entdeckung und Nutzung) den

Beurteiler:innen zuzuordnen (Funder, 1995). Die eigentliche Persönlichkeitsbeurteilung erfolgt erst im letzten Schritt, da erst hier Rückschlüsse auf die Persönlichkeitseigenschaft der Zielperson gezogen werden (Funder, 1995), sodass diese Prozessvariable für das Ziel der vorliegenden Arbeit von besonderem Interesse ist. Nach Back und Nestler (2016) besteht allerdings eine große Herausforderung darin, diese beiden letzten Schritte des RAM empirisch voneinander zu trennen, was auch bereits von Funder (1999) angemerkt wurde.

Abbildung 2.2

Erweiterte Abbildung des Realistic Accuracy Model von Funder (1995)



Anmerkung. Die Abbildung wurde in Anlehnung an die Darstellung bei Funder (2012) erstellt.

Die Aufteilung in die vier Schritte leitete Funder (1995) aus der Grundannahme des RAM ab, dass Persönlichkeitseigenschaften „real characteristics of individuals" (S. 653) sind und akkurate Urteile sich daher aus Eigenschaften der beurteilenden Person und der Zielperson ergeben. Später ergänzte Funder (1999) zudem, dass es sich bei den vier Schritten um Aufspaltungen der Cue Validity und Cue Utilization aus dem Linsenmodell (Brunswik, 1956) in jeweils zwei Schritte handelt. Darüber hinaus entspricht die Übereinstimmung zwischen Persönlichkeitseigenschaft und Beurteilung laut Funder (1999, 2012) dem, was Brunswik als Functional Validity oder (Perceptual) Achievement bezeichnete.

Neben den vier Schritten werden im RAM auch vier Moderatorvariablen unterschieden, die sich auf Grund ihres Einflusses in mindestens einem der vier Schritte auf die Akkuratheit der Beurteilungen auswirken sollen: *Good Target*, *Good Trait*, *Good Information* und *Good Judge* (Funder, 1995). Zudem schlug Funder (1995) mögliche Einflüsse von zweifach-Interaktionen der vier Moderatoren vor. Die ersten drei der genannten Moderatoren beziehen sich auf die Annahme des Modells, dass es individuelle Unterschiede darin gibt, wie einfach bei bestimmten Zielpersonen (Good Target) und bestimmten Persönlichkeitseigenschaften (Good Trait) eine akkurate Beurteilung vorgenommen werden kann sowie dass es Information

gibt, die eine akkuratere Beurteilung erleichtert (Good Information; Funder, 1995). Von besonderem Interesse ist der vierte Moderator (Good Judge), der sich laut Funder auf interindividuelle Unterschiede in den letzten beiden Schritten des RAM bezieht und in mehrere Komponenten aufgeteilt werden kann.

Als eine der Komponenten des Moderators Good Judge nannte Funder (1995) eine Fähigkeit zur validen Nutzung der entdeckten Hinweise, für die er auch die Bezeichnung *Judgmental Ability* verwendete und die sich auf interindividuelle Unterschiede im letzten Schritt des RAM (Nutzung) bezieht. Funder vermutete unter Verweis auf zwei Studienergebnisse zudem Zusammenhänge der Judgmental Ability mit der allgemeinen Intelligenz. Eine Definition oder nähere Beschreibung dieser Fähigkeit fehlt allerdings sowohl in der ersten Publikation zum RAM (Funder, 1995) als auch in neueren Veröffentlichungen (z.B. Letzring & Funder, 2021). Bei Letzring und Funder (2021) werden allerdings mehrere spezifischere Teilfähigkeiten aufgezählt, die im letzten Schritt des RAM eine Rolle spielen sollen. Diese spezifischen Teilfähigkeiten können somit am ehesten als etwas umfassendere Charakterisierung der Judgmental Ability angesehen werden. Nach Letzring und Funder sind dies:

1. Die Fähigkeit, über die Relevanz der Hinweise zu entscheiden
2. Die Fähigkeit, die Hinweise angemessen zu gewichten
3. Die Fähigkeit, die Hinweise zu kombinieren
4. Die Fähigkeit, andere mögliche Ursachen für das Verhalten in Betracht zu ziehen (z.B. situative Einflüsse).

Allerdings bleibt bei es bei Letzring und Funder (2021) bei einer Aufzählung der spezifischen Teilfähigkeiten ohne Annahmen dazu, was diese gemeinsam haben und somit, was im Kern genau unter einer Fähigkeit zur validen Nutzung der Hinweise verstanden werden kann. Des Weiteren fehlt eine explizite Begründung der vorgeschlagenen Teilfähigkeiten. Deren Beschreibungen lassen jedoch vermuten, dass sie aus den Ursprüngen des RAM abgeleitet wurden (insb. Attributionstheorien sowie Linsenmodell; vgl. Letzring & Funder, 2021). Neben einer Fähigkeit zur validen Nutzung nannte Funder (1995, 1999) zudem eine Fähigkeit zur korrekten Wahrnehmung von Hinweisen, die sich auf den vorletzten Schritt des RAM (Entdeckung) auswirken soll. Der eigentliche Vorgang der Persönlichkeitsbeurteilung ist hiervon also nicht betroffen, sodass diese Fähigkeit nicht näher betrachtet wird.

Eine weitere angenommene Komponente des Moderators Good Judge ist Wissen über Persönlichkeit und wie sich Persönlichkeit im Verhalten manifestiert (Funder, 1995). Laut Funder (1995) soll das Wissen beispielsweise das Erkennen von Hinweisen erleichtern.

Demzufolge kann diese Komponente erneut am ehesten dem vorletzten Schritt des RAM (Entdeckung) zugeordnet werden. Bei Letzring und Funder (2021) sowie in anderen Publikationen, die sich auf das RAM beziehen (Christiansen et al., 2005; De Kock et al., 2020), wird persönlichkeitsrelevantes Wissen wiederum dem letzten Schritt (Nutzung) zugeordnet. Die Wissenskomponente wurde später zudem von Christiansen et al. (2005) aufgegriffen und in Form der Dispositionellen Intelligenz theoretisch näher beleuchtet (vgl. Abschnitt 2.2.5).

Die letzte Komponente des Moderators Good Judge ist die Motivation der Beurteiler:innen, unter anderem eine akkurate Beurteilung vorzunehmen (Funder, 1995). Diese Komponente ist im Hinblick auf die nähere Beschreibung einer Fähigkeit der Beurteiler:innen nicht von Interesse und wird daher nicht weiter betrachtet.

Gedacht war das RAM laut Funder (1995) vor allem als theoretischer Orientierungsrahmen zur Organisation der bisherigen und zukünftigen Forschung zu akkuraten Persönlichkeitsurteilen. Grundlage des Modells waren unter anderem das Linsenmodell von Brunswik (1956) sowie die von Funder (1995) bereits zu einem früheren Zeitpunkt vorgenommene Organisation bisheriger Befunde zu Einflussfaktoren auf akkurate Persönlichkeitsurteile in die vier Bereiche Good Target, Good Trait, Good Information und Good Judge (Funder, 1995). Der Autor sah das RAM als nächsten Schritt an, um zu erklären, wie akkurate Beurteilungen zustande kommen sowie um den Einfluss der Moderatoren auf die Akkuratheit zu erklären. Er bezeichnete das RAM dabei als initial und prototypisch, was damit zusammenpasst, dass Funder bei der ersten Vorstellung seines Modells im Jahr 1995 für einige Bestandteile und Annahmen seines Modells nur wenig empirische Befunde anführte oder anführen konnte. Insbesondere die präsentierte Evidenz zum Moderator Good Judge, was Funder (1995) selbst anmerkt (vgl. auch Funder, 1999), sowie zu den angenommenen Interaktionen der Moderatoren ist eher gering. Letzring und Funder (2021) weisen im Zusammenhang mit den Interaktionen ebenfalls darauf hin, dass diese auf theoretischen Überlegungen basieren und bisher nicht systematisch überprüft wurden.

Betrachtet man die Forschung zur Akkuratheit von Persönlichkeitsbeurteilungen, wird deutlich, dass mittlerweile viele Studien unter Verweis auf das RAM im Ganzen oder auf einen bestimmten Aspekt des RAM durchgeführt werden (z.B. Chen et al., 2018, Darbyshire et al., 2016; Tackett, 2011). Die Relevanz des Modells für das Forschungsgebiet kann somit als hoch und Funders (1995) Ziel, einen Orientierungsrahmen zur Verfügung zu stellen, als – zumindest zu einem gewissen Teil – erreicht angesehen werden. Auch für das Ziel der vorliegenden Arbeit schafft das RAM einen sinnvollen theoretischen Orientierungsrahmen. Die explizite Nennung

und Einordnung einer Fähigkeit der Beurteiler:innen zur valideren Nutzung entdeckter Hinweise unterstreicht zudem den Nutzen des Modells.

2.2.3 State and Trait Accuracy Model

Ein weiteres Modell, das unter anderem auf dem RAM von Funder (1995) aufbaut und somit als Erweiterung angesehen werden kann, ist das State and Trait Accuracy Model (STAM) von Hall et al. (2017). Das STAM wurde als Zwei-Phasen-Modell konzipiert, das eine kausale Beziehung zwischen akkuraten Urteilen über affektive Zustände und akkuraten Urteilen über Persönlichkeitseigenschaften annimmt (Hall et al., 2017). In Phase 1 des Modells erfolgt zunächst auf Basis behavioraler Information ein Urteil über die affektiven Zustände einer Zielperson. Diese Information wird in Phase 2 wiederum genutzt, gegebenenfalls zusammen mit weiteren Hinweisen, um ein Urteil über die Persönlichkeitseigenschaft der Zielperson abzugeben (Hall et al., 2017). Wie die Autor:innen des Modells erläutern, soll jede der beiden Phasen wiederum nach den Schritten des RAM ablaufen und zudem jeweils im letzten Schritt kontextuelle Information mit in das Urteil einbezogen werden. Neben dem RAM basiert das STAM auf Attributionsmodellen von Trope und Gilbert (Hall et al., 2017). Zudem erläutern Hall et al. Gründe für die Richtung der kausalen Beziehung von Zuständen zu Persönlichkeit, allerdings ohne diese mit entsprechenden Quellen zu belegen. Empirische Evidenz zur Unterstützung des Modells ist bisher auf korrelative Befunde zum Zusammenhang zwischen akkuraten Urteilen über Neurotizismus und negativem Affekt sowie – zu einem geringeren Ausmaß – zwischen akkuraten Urteilen über Extraversion und positivem Affekt beschränkt, wobei diese Evidenz von den Autor:innen des Modells selbst stammt (Hall et al., 2017). Weitere Studien zur Überprüfung der Annahmen des Modells sind zwar derzeit in Vorbereitung (Letzring & Funder, 2018), wurden bisher allerdings nur als Posterpräsentationen veröffentlicht (Letzring & Funder, 2021). Das Modell muss auf Grund der teilweise unklaren theoretischen Herleitung sowie (noch) geringen empirischen Evidenz somit als vorläufig betrachtet werden.

2.2.4 Soziale Intelligenz

Ein weiterer Forschungsbereich, der sich unter anderem auch mit der Akkuratheit von Persönlichkeitsbeurteilungen beschäftigt und relevante Aspekte hinsichtlich einer Fähigkeit der Beurteiler:innen thematisiert, ist der zur *Sozialen Intelligenz* (SI). Dies wird bereits bei der Betrachtung verschiedener Definitionen des Konstruktes deutlich. Im Jahr 1933 definierte Vernon SI als

> ability to get along with people in general, social technique or ease in society, knowledge of social matters, susceptibility to stimuli from other members of a group, as well as insight into the temporary moods or the underlying personality traits of friends and of strangers. (S. 44)

Dieser Definition zufolge kann das Verständnis von Persönlichkeitseigenschaften somit als eine Teilfähigkeit der SI angesehen werden. In der zitierten Arbeit beschäftigte sich Vernon (1933) zudem nicht primär mit der SI, sondern mit interindividuellen Unterschieden in einer Fähigkeit zur korrekten Beurteilung der Persönlichkeit. In seiner Studie setzte er eine Reihe unterschiedlicher Tests ein, bei denen unter anderem prominente Personen anhand von Fotos hinsichtlich verschiedener Persönlichkeitseigenschaften in eine Rangfolge gebracht werden mussten. In einem anderen Test mussten die Testpersonen zudem fünf auf Fotos dargestellten, fremden Zielpersonen vorgegebene Persönlichkeitseigenschaften zuordnen (Vernon, 1933). Vernon konnte in seiner Studie zeigen, dass eine bessere Leistung in der Beurteilung der Persönlichkeit von Fremden mit höherer sozialer (r = .36) sowie allgemeiner Intelligenz (r = .31) zusammenhängt. Hierbei muss allerdings beachtet werden, dass seine gewählten Operationalisierungen der Fähigkeit zur korrekten Beurteilung der Persönlichkeit Einschränkungen bezüglich der Konstruktvalidität aufwiesen (u.a. Zusammenfassung von Tests trotz fehlendem sinnvollem Interkorrelationsmuster; Vernon, 1933).

Wedeck (1947) beschäftigte sich mit der sogenannten *Psychological Ability*, die er als „ability to judge correctly the feelings, moods, motivations of individuals" (S. 133) beschrieb und die in neuerer Literatur der SI zugeordnet wird (O'Sullivan et al., 1965; Weis & Süß, 2005). Wie in Abschnitt 2.1 beschrieben wurde, können Gefühle und Motivation als Bestandteil der Persönlichkeit angesehen werden (McCrae & Costa, 2003, Mischel & Shoda, 1995), sodass Überschneidungen zwischen der von Wedeck beschriebenen Fähigkeit und dem akkuraten Beurteilen von Persönlichkeitseigenschaften vorliegen. Wedeck selbst nahm an, dass die Psychological Ability eine Fähigkeit darstellt, die benötigt wird, um die Persönlichkeit anderer Personen korrekt beurteilen zu können. Er schlug zudem eine Reihe an Tests vor, um die Fähigkeit zu operationalisieren, die in seinen Analysen zumindest zum Teil Primärladungen auf einem gemeinsamen Faktor aufwiesen. Hierzu gehörte beispielsweise ein Test, bei dem Portraitfotos von Zielpersonen präsentiert wurden, die hinsichtlich verschiedener Tendenzen (z.B. Hartherzigkeit, Selbstbewusstsein) beurteilt werden mussten (Wedeck, 1947).

Die Definition eines wichtigen Teilbereichs der SI, der sogenannten *Behavioralen Kognition* (Behavioral Cognition), nahmen O'Sullivan und Guilford (1975) vor. Der Autorin und dem Autor zufolge handelt es sich bei diesem Teilbereich um die „ability to understand

the thoughts, feelings, and intentions of other people insofar as they are manifested in discernible behavior" (O'Sullivan & Guilford, 1975, S. 256; vgl. auch O'Sullivan et al., 1965). Neben Gedanken, Gefühlen und Intentionen gehören laut O'Sullivan et al. (1965) auch andere psychologische Dispositionen, die das soziale Verhalten von Personen beeinflussen können, zum Inhaltsbereich der Behavioralen Kognition (vgl. Abschnitt 2.2.4.1). Und auch wenn Persönlichkeitsbeurteilungen hier nicht explizit genannt werden, werden in der Beschreibung dieses Teilbereichs der SI ebenfalls Gemeinsamkeiten mit der akkuraten Beurteilung von Persönlichkeitseigenschaften deutlich, da mit Gedanken, Gefühlen und (Ursachen für) Verhalten wichtige Bestandteile der Persönlichkeit angesprochen werden (McCrae & Costa, 2003; vgl. Abschnitt 2.1).

2.2.4.1 Structure-of-Intellect Model. Die zuvor beschriebene Behaviorale Kognition umfasst sechs von insgesamt 30 Teilfähigkeiten, die dem sogenannten Structure-of-Intellect (SOI) Model von Guilford (1967, 1988) zufolge der SI zugeordnet werden können (O'Sullivan & Guilford, 1975; O'Sullivan et al., 1965). Beim SOI-Model handelt es sich um ein Intelligenzstrukturmodell, in dem anfänglich 120 intellektuelle Fähigkeiten beschrieben wurden (Guilford, 1967) und das in der aktuellsten Version auf 180 intellektuelle Fähigkeiten erweitert wurde (Guilford, 1988). Diese Fähigkeiten ergeben sich dem Modell nach aus der Kombination der drei Facetten Operation (beteiligter mentaler Prozess), Inhalt (Art der vorhandenen Information) und Produkt (Form, in der die vorhandene Information erfasst wird bzw. Ergebnis der mentalen Verarbeitung; Guilford, 1967, 1985; O'Sullivan et al., 1965). Es werden in der aktuellsten Version sechs Operationen, fünf Inhalte sowie sechs Produkte unterschieden, durch deren vollständige Kombination sich die 180 Fähigkeiten ergeben (Guilford, 1988). Von besonderem Interesse ist der Inhalt *behavioral*, der Information umfasst, die sich auf das Verhalten anderer Personen bezieht (Guilford, 1967). Zu dieser behavioralen Information gehören laut Guilford (1967) und O'Sullivan et al. (1965) Gefühle, Gedanken, Intentionen, Motive, Einstellungen und andere psychologische Dispositionen, die sich im (sozialen) Verhalten einer Person zeigen. Im SOI-Model wird der gesamte behaviorale Inhaltsbereich der SI zugeordnet und hier insgesamt 30 Teilfähigkeiten angenommen, die sich durch die Kreuzung von fünf Operationen mit den sechs Produkten innerhalb des behavioralen Inhalts ergeben (Guilford, 1985; O'Sullivan & Guilford, 1975; O'Sullivan et al., 1965).[1] Die im vorherigen Abschnitt zitierte Definition von O'Sullivan und Guilford (1975) des SI-

[1] Die Unterscheidung von sechs Operationen wurde erst nach Publikation der Arbeiten zum behavioralen Inhaltsbereich vorgeschlagen (Guilford, 1988).

Teilbereichs bezieht sich nur auf eine der fünf Operationen, die als *Kognition* bezeichnet wurde und die die Entdeckung, das Wissen oder Verstehen von Information beziehungsweise das Strukturieren von Information umfasst (Guilford, 1985). Im Hinblick auf eine Fähigkeit, akkurate Persönlichkeitsbeurteilungen vornehmen zu können, erscheinen neben der Kognition zwei weitere der Operationen potentiell relevant. Dies ist zum einen die Operation *konvergente Produktion*, die das Abrufen einer bestimmten, präzisen Information aus dem Gedächtnis beschreibt und unter anderem das Ziehen einer korrekten Schlussfolgerung aus vorgegebenen Informationen umfasst (Guilford, 1985) beziehungsweise das Generieren einer korrekten Lösung (O'Sullivan et al., 1965). Bezogen auf den behavioralen Inhaltsbereich könnte es sich hierbei um Schlussfolgerungen über die Gedanken, Gefühle oder Intentionen und somit auch die Persönlichkeitseigenschaften einer Zielperson handeln, die auf Basis vorhandener behavioraler Information vorgenommen werden. O'Sullivan et al. (1965) nennen als Beispiel allerdings die Fähigkeit „doing just the right thing at the right time" (S. 5), was bedeutet, dass es sich hierbei eher um eine Fähigkeit mit Verhaltenskomponente handeln soll, die für akkurate Persönlichkeitsbeurteilungen keine Rolle spielt. Auch O'Sullivan und Guilford (1975) beschreiben die behaviorale konvergente Produktion als Fähigkeit, korrekte soziale Reaktionen zu zeigen. Leider finden sich zu dieser spezifischen Fähigkeit keine weiteren Beschreibungen oder Veröffentlichungen, sodass offen bleibt, was genau hierunter verstanden werden kann. Die zweite potentiell interessante Operation *Evaluation* beschreibt den Entscheidungsprozess, ob und gegebenenfalls wie gut ein Item logischen Voraussetzungen entspricht – beispielsweise, ob ein Gegenstand bestimmte Eigenschaften besitzt (Guilford, 1985). Nach O'Sullivan et al. (1965) kann die behaviorale Evaluation beschrieben werden als Fähigkeit, das Verhalten anderer Personen einzuschätzen, beispielsweise anhand des Kriteriums Angemessenheit. Ob eine Einschätzung des Verhaltens anhand der zugrundeliegenden Gedanken, Gefühle oder Intentionen beziehungsweise Persönlichkeitseigenschaften ebenfalls hierzu gezählt werden kann, bleibt allerdings offen, da auch hier keine weiteren Beschreibungen oder Veröffentlichungen vorliegen. O'Sullivan und Guilford (1975) beschreiben diese Teilfähigkeit wiederum als Fähigkeit, vernünftige soziale Entscheidungen zu treffen, was einen anderen Schwerpunkt setzt. Der SOI-Teilbereich Behaviorale Kognition wurde hingegen deutlich umfangreicher untersucht. O'Sullivan et al. (1965) nahmen eine Operationalisierung dieses Fähigkeitsbereichs anhand mehrerer Leistungstestaufgaben vor, in denen zumindest vereinzelt auch Persönlichkeitseigenschaften in den Aufgabeninhalten wiederzufinden sind (vgl. auch O'Sullivan & Guilford, 1975). So werden in der von den Autor:innen verwendeten Aufgabe Odd Strip Out drei kurze Cartoons präsentiert, von denen derjenigen Cartoon ausgewählt

werden soll, in dem die Hauptfigur anders reagiert als in den anderen beiden. Den Autor:innen zufolge erfordert diese Aufgabe einen Rückschluss von den präsentierten Reaktionen auf die zugrundliegende Persönlichkeitseigenschaft der Hauptfigur (O'Sullivan & Guilford, 1975; O'Sullivan et al., 1965). In einer weiteren Aufgabe mit dem Namen Cartoon Predictions wird ebenfalls zunächst ein kurzer Cartoon präsentiert. Die Testpersonen werden im Anschluss dazu aufgefordert, das Verhalten der gezeigten Hauptfigur vorherzusagen, indem sie den Cartoon durch Auswahl eines von drei Bildern fortsetzen. Laut O'Sullivan et al. (1965) bestand das ursprüngliche Ziel darin, individuelle Persönlichkeiten in den Cartoons darzustellen und deren individuelles Verhalten vorhersagen zu lassen. Hierfür sei aber die Präsentation umfangreicher Information über die Hauptfigur notwendig gewesen, was zu unökonomischen Aufgaben geführt hätte. Daher wurde dieses Ziel verworfen und Cartoons konstruiert, die Vorhersagen von stereotypischen Verhaltensweisen erfordern (O'Sullivan et al., 1965). Von besonderem Interesse ist zudem, dass die Autor:innen darauf hinwiesen, dass die Erfassung der wahrgenommenen Persönlichkeit anderer Personen auf Basis deren Verhaltensweisen in unterschiedlichen Situationen eine geeignete Operationalisierung des Teilbereichs Behaviorale Kognition darstellt und nur auf Grund von Schwierigkeiten in der Testkonstruktion nicht umgesetzt wurde. Insgesamt konnten O'Sullivan et al. (1965; vgl. auch O'Sullivan & Guilford, 1975) in ihrer Studie die angenommene faktorielle Struktur der neu konstruierten Tests überwiegend bestätigen und den untersuchten Teilbereich der SI von unter anderem verbalen und räumlichen Fähigkeiten abgrenzen (siehe aber auch Riggio et al., 1991; Romney & Pyryt, 1999).

2.2.4.2 Integratives Modell Sozialer Intelligenz. Das Integrative Modell Sozialer Intelligenz von Weis und Süß (2005; vgl. auch Weis et al., 2006) wurde dem Namen entsprechend mit dem Ziel entwickelt, frühere Forschung zur SI zu integrieren und ermöglicht damit die Einordnung verschiedener Konzeptualisierungen. So kann die zitierte Definition der SI von Wedeck (1947) sowie der SI-Teilbereich Behaviorale Kognition aus dem SOI-Model (O'Sullivan & Guilford, 1975; O'Sullivan et al., 1965) dem Modell zufolge einer bestimmten Teilfähigkeit der SI zugeordnet werden, die als Soziales Verständnis bezeichnet wird (Weis & Süß, 2005). Auch der letzte Teil der SI-Definition von Vernon (1933), der sich auf das Verständnis von Stimmungen und Persönlichkeitseigenschaften bezieht, lässt sich hier einordnen (Seidel, 2007; Weis, 2008). Wie Weis und Süß (2005) erläutern, können Konzeptualisierungen der SI danach unterschieden werden, ob sie sich auf eine kognitive Komponente (z.B. Verstehen, Beurteilen) oder eine Verhaltenskomponente (z.B. vernünftig

handeln) der SI fokussieren. Beispielsweise umfasst die Definition von Vernon (1933) beide Komponenten, die Definition der Behavioralen Kognition von O'Sullivan und Guilford (1975) hingegen nur die kognitive Komponente. Der primäre Fokus des Integrativen Modells liegt dieser Unterteilung nach ebenfalls auf der kognitiven Komponente (Weis & Süß, 2005).

Dem Integrativen Modell Sozialer Intelligenz zufolge handelt es sich bei der SI um eine Fähigkeit, die aus fünf Teilfähigkeiten besteht (Weis & Süß, 2005):

1. Soziales Verständnis: Hierbei handelt es sich laut Weis und Süß (2005) um das zentrale Konstrukt der bisherigen Forschung. Soziales Verständnis kann definiert werden als „die Fähigkeit, gegebene soziale Informationen in der jeweiligen Situation zu verstehen und korrekt zu interpretieren" (Weis et al., 2006, S. 224).
2. Soziales Gedächtnis: Diese Teilfähigkeit umfasst „intentionales Speichern und Abrufen von unterschiedlich komplexen, episodischen und semantischen Gedächtnisinhalten einer sozialen Situation" (Weis et al., 2006, S. 224).
3. Soziale Wahrnehmung: Hierbei handelt es sich um „die Fähigkeit, soziale Informationen (möglichst schnell) wahrzunehmen" (Weis et al., 2006, S. 224).
4. Soziale Kreativität oder Flexibilität: Bei dieser Teilfähigkeit geht es darum, möglichst viele verschiedene „Interpretationen oder Lösungen für eine gegebene soziale Situation" (S. 224) zu produzieren (Weis et al., 2006).
5. Soziales Wissen: Diese Teilfähigkeit beschreibt prozedurales Wissen über soziale Situationen und Sachverhalte (Weis & Süß, 2005; Weis et al., 2006).

Für die vorliegende Arbeit von besonderem Interesse ist die Teilfähigkeit Soziales Verständnis beziehungsweise *Social Understanding* (SU). Dies wird allerdings erst bei einer näheren Betrachtung des Konstrukts deutlich, da auf Basis der oben zitierten Definition von SU nur schwer ersichtlich ist, welche Leistungen dieser Fähigkeit zugeordnet werden können. Zudem kann die Definition als zum Teil zirkulär bezeichnet werden. Weis und Süß (2005) präzisieren die Definition anhand des Beispiels, dass das korrekte Verstehen der (non)verbalen Kommunikation einer Person der SU zugeordnet werden kann. Zudem wird ergänzt, dass die zu verstehende soziale Information unterschiedlich komplex sein kann, unterschiedliche situative Auswirkungen haben kann sowie Hinweise auf verschiedene zugrunde liegende Merkmale darstellen kann (Weis et al., 2006; Weis & Süß, 2005). Weitere Präzisierungen werden hier allerdings nicht vorgenommen. Eine informativere Präzisierung findet sich in der Dissertation von Weis (2008; vgl. auch Seidel, 2007). Die Autorin weist hier zum einen darauf hin, dass die zentrale kognitive Anforderung von SU logisches Schlussfolgern (Reasoning) darstellt und hierzu auch die kognitiven Operationen Verstehen, Interpretieren, Beurteilen,

Einsicht, Vorhersagen und Nachvollziehen gezählt werden können (Weis, 2008). Zum anderen konkretisiert Weis, dass die soziale Information Schlussfolgerungen über die Emotionen, Gedanken, Intentionen, Motivationen und Persönlichkeitseigenschaften einer Person erlauben sollen. Hier wird schließlich die Relevanz von SU für die vorliegende Arbeit deutlich, da dieser Konzeptualisierung von SU zufolge das Schlussfolgern über Persönlichkeitseigenschaften anderer Personen auf Basis vorliegender Information als Teilfähigkeit oder Teilleistung von SU angesehen werden kann. Darüber hinaus vergleicht Weis die Definition von SU explizit mit dem Konzept der akkuraten Beurteilung von Personen, wozu auch die akkurate Persönlichkeitsbeurteilung gehört.

Basierend auf dem Integrativen Modell Sozialer Intelligenz wurde ein Test zur Erfassung der postulierten Teilfähigkeiten konstruiert – der Magdeburger Test zur Sozialen Intelligenz (MTSI; Conzelmann et al., 2013; Süß et al., 2009). Bei den MTSI-Aufgaben zur Erfassung von SU handelt es sich um mehrere Szenarien, in denen Information über jeweils eine reale Zielperson in Form von Text, Audio, Bildern oder Videos ohne Ton präsentiert wird. Nach jeder gegebenen Information werden die Testpersonen gebeten, die Emotionen, Kognitionen und Beziehungen der jeweiligen Zielperson zu anderen Personen anhand einer 7-stufigen Likert-Skala zu bewerten (Conzelmann et al., 2013; Seidel, 2007; Süß et al., 2009; Weis, 2008). Am Ende eines SU-Szenarios müssen die Testpersonen zudem die Persönlichkeit der Zielpersonen beurteilen, unter anderem anhand der Big Five-Persönlichkeitsfaktoren (Seidel, 2007; Süß et al., 2009; Weis, 2008). Die Persönlichkeitsbeurteilungen stellen allerdings nur einen verhältnismäßig kleinen Teil der Items innerhalb eines Szenarios dar, auch da jeder der Big Five-Faktoren nur anhand eines einzelnen Items eingeschätzt werden muss (Süß et al., 2009). Zudem werden diese Persönlichkeitsbeurteilungen bei der Auswertung der SU-Aufgaben normalerweise nicht in die Bildung eines Gesamtwertes für SU miteinbezogen (Süß et al., 2009). Warum die Persönlichkeitsbeurteilungen beim SU-Gesamtwert außen vor gelassen werden, bleibt allerdings unklar. Eine mögliche Erklärung liefert das Ergebnis einer Analyse von Seidel (2007). Im Rahmen einer konfirmatorischen Faktorenanalyse, in dem der Zusammenhang zwischen SU und Sozialem Gedächtnis geschätzt wurde, zeigte sich ein schlechterer Modellfit, wenn die Persönlichkeitsbeurteilungen in das Modell miteinbezogen wurden (Seidel, 2007). Seidel vermutete auf Basis dieser Ergebnisse, dass die Persönlichkeitsbeurteilungen möglicherweise doch andere Aspekte erfassen als die anderen Items der SU-Aufgaben. In jedem Fall lässt sich daraus schließen, dass eine Fähigkeit, akkurate Persönlichkeitsbeurteilungen vornehmen zu können, im MTSI nicht angemessen abgebildet wird. Abgesehen davon konnte gezeigt werden, dass die MTSI-Aufgaben zur Erfassung von

SU einen eigenständigen Faktor bilden, der sich von den klassischen Intelligenzkonstrukten logisches Schlussfolgern, Merkfähigkeit und Bearbeitungsgeschwindigkeit empirisch abgrenzen lässt (Conzelmann et al., 2013). Allerdings konnten Conzelmann et al. (2013) entgegen der Erwartungen nur eine maximal mittlere latente Korrelation zu anderen Teilfähigkeiten der SI aufzeigen.

2.2.4.3 Theorie der Multiplen Intelligenzen. Auch Gardners (1983, 1991) Theorie der Multiplen Intelligenzen beinhaltet eine Fähigkeit, die inhaltliche Überschneidungen mit der SI aufweist. Gardner war der Meinung, dass die klassischen Intelligenztheorien und -tests zu eng gefasst sind und nicht alle relevanten Bereiche der menschlichen Intelligenz berücksichtigen (Gardner, 1991; Gardner & Hatch, 1989). Er wollte daher eine konkurrierende, neue Sichtweise auf die Intelligenz formulieren und orientierte sich bei der Zusammenstellung der seiner Ansicht nach relevanten Fähigkeiten an acht Kriterien, wie beispielsweise dem Ausfall einzelner Fähigkeiten nach Verletzungen des Gehirns, überdurchschnittliche Ausprägung bestimmter Fähigkeiten bei Kindern, Feststellung einer zentralen kognitiven Operation oder die Identifikation der normalen Entwicklung einer Fähigkeit (Gardner, 1991). Auf Grundlage dessen schlug Gardner (1991) zunächst sieben Intelligenzen vor, die später um weitere ergänzt wurden (vgl. Gardner & Moran, 2006): Linguistische, musikalische, logisch-mathematische, räumliche, körperlich-kinästhetische, intrapersonale und interpersonale Intelligenz. Von potentiellem Interesse sind die beiden letztgenannten personalen Intelligenzen. Die *interpersonale Intelligenz* wurde beschrieben als Fähigkeit, zwischen den Stimmungen, Temperamenten, Motivationen und Bedürfnissen anderer Personen zu unterscheiden und angemessen auf diese zu reagieren (Gardner, 1991; Gardner & Hatch, 1989). Im Gegensatz dazu verstand Gardner (1991) unter der *intrapersonalen Intelligenz* den „Zugang zum eigenen Gefühlsleben“ (S. 219), wozu er unter anderem die Fähigkeit zählte, zwischen eigenen Gefühlen zu differenzieren, die eigenen Gefühle zu benennen, das eigene Verhalten mit Hilfe der Gefühle zu steuern sowie Wissen über sich selbst (Gardner, 1991; Gardner & Hatch, 1989). Auf Basis dieser Beschreibungen werden Überschneidungen der *inter*personalen Intelligenz mit weiter oben zitieren Definitionen der SI deutlich (Matthews et al., 2012). Auch im Hinblick auf eine Fähigkeit, die bei akkuraten Persönlichkeitsbeurteilungen eine zentrale Rolle spielt, erscheint die interpersonale Intelligenz mit ihrem Fokus auf persönlichkeitsrelevante Aspekte anderer Personen interessant. Allerdings muss hierbei beachtet werden, dass Gardners Theorie in der Vergangenheit vielfach kritisiert wurde. Die Kritik bezog sich unter anderem auf die fehlende theoretische Fundierung und mangelnde empirischer Evidenz (Klein, 1997, 1998;

Waterhouse, 2006a, 2006b), sodass auch das Konzept der interpersonalen Intelligenz mit Vorsicht betrachtet werden muss und als theoretische Grundlage weniger geeignet erscheint.

2.2.4.4 Zusammenfassende Bewertung der Theorien und Modelle. Insgesamt betrachtet kann die SI im Hinblick auf das Ziel der Konzeptualisierung einer Fähigkeit zur akkuraten Beurteilung der Persönlichkeit anderer Personen als wichtige und relevante theoretische Grundlage betrachtet werden. Bereits die frühen Definitionen der SI von Vernon (1933) und Wedeck (1947) sowie deren gewählte Operationalisierungen von SI, die unter anderem Persönlichkeitsbeurteilungen umfassten, machen die Relevanz deutlich. Gleiches gilt für die Beschreibung und vorgeschlagene Erfassung des Teilbereichs der Behavioralen Kognition aus dem SOI-Model (O'Sullivan & Guilford, 1975; O'Sullivan et al., 1965). Den größten Nutzen für die vorliegende Arbeit hat die Teilfähigkeit SU aus dem Integrativen Modell Sozialer Intelligenz von Weis und Süß (2005), zusammen mit den ergänzenden Ausführungen von Weis (2008; vgl. auch Seidel, 2007). Durch die Annahme, dass sich SU auch auf Persönlichkeitseigenschaften anderer Personen bezieht (Weis, 2008), kann bei der Konzeptualisierung einer Fähigkeit im Bereich akkurater Persönlichkeitseigenschaften auf die Beschreibung der zentralen kognitiven Anforderung von SU zurückgegriffen werden. Von Vorteil ist hier zudem, dass den Autor:innen des Integrativen Modells zufolge die anderen relevanten Definitionen der SI von Vernon (1933), Wedeck (1947) sowie O'Sullivan und Guilford (1975) ebenfalls der SU zugeordnet (Seidel, 2007; Weis, 2008; Weis & Süß, 2005) und somit als zusätzlicher theoretischer Rahmen angesehen werden können. Eine noch zu klärende Frage ist allerdings, inwieweit eine inhaltliche Abgrenzung von SU und einer Fähigkeit zur akkuraten Persönlichkeitsbeurteilung möglich ist oder ob letztere lediglich eine Teilfähigkeit von SU darstellt (vgl. Abschnitt 3.1.1). Zudem stellt die Theorie der Multiplen Intelligenzen von Gardner (1983, 1991) auf Grund der vorhandenen Kritik keine gute Grundlage dar und sollte daher nicht weiter berücksichtigt werden.

2.2.5 Dispositionelle Intelligenz

Christiansen et al. (2005) führten das Konstrukt der *Dispositionellen Intelligenz* (Dispositional Intelligence; DI) ein und definierten es als „knowledge about personality and how it is related to behavior" (S. 124). DI umfasst drei deklarative Wissensstrukturen, die als Komponenten der DI aufgefasst werden können und von Christiansen et al. folgendermaßen beschrieben wurden (für die Namen der Komponenten siehe De Kock et al., 2015):

1. Trait Induction: Wissen über die Verbindung zwischen Persönlichkeitseigenschaft und Verhalten. Hierzu gehört die Fähigkeit, zu erkennen, auf welche Persönlichkeitseigenschaft(en) ein Verhalten Rückschlüsse erlaubt.
2. Trait Contextualization: Verständnis der Relevanz von Situationen für Persönlichkeitseigenschaften. Hierzu gehört das Wissen, in welchen Situationen sich Persönlichkeitseigenschaften im Verhalten zeigen und wann Verhalten hauptsächlich situativ bedingt ist.
3. Trait Extrapolation: Kenntnis von Konzepten über Persönlichkeitseigenschaften. Hierzu gehört das Verständnis, welche Persönlichkeitseigenschaften normalerweise gemeinsam bei einer Person auftreten.

Ausgangspunkt für die DI waren Überlegungen von Christiansen et al. (2005) über den Prozess akkurater Persönlichkeitsbeurteilungen sowie, dass es sich hierbei um einen komplexen kognitiven Prozess handelt, an dem neben allgemeiner Intelligenz auch weitere spezifischere Fähigkeiten beteiligt sein müssten. Bei der Ableitung der DI und der Begründung deren Komponenten bezogen sich Christiansen et al. zudem auf das RAM von Funder (1995; zitiert wird allerdings ausschließlich das Buch aus dem Jahr 1999) und die dort postulierte Wissenskomponente (vgl. Abschnitt 2.2.2) sowie ein Modell aus dem Bereich der Attributionstheorien (Trope, 1986). Christiansen et al. zufolge spielt DI beim Schlussfolgern über dispositionelle Information eine wichtige Rolle und kann verglichen werden mit Wissen über Wörter und Grammatik und dessen Rolle beim verbalen Schlussfolgern.

Bei DI handelt es sich somit um ein Wissenskonstrukt. Dies zeigt sich auch in der von Christiansen et al. (2005) vorgeschlagenen Operationalisierung durch einen neu entwickelten Leistungstest, der sich aus 45 Multiple-Choice Fragen zu den drei Komponenten der DI zusammensetzt. Die Fragen beziehen sich auf etablierte Forschungsergebnisse aus der Persönlichkeitspsychologie, wobei diejenige Antwortoption als richtig gewertet wird, die am ehesten mit den Forschungsergebnissen sowie theoretischen Annahmen übereinstimmt (Christiansen et al., 2005). Die Testinhalte sind also nicht die Persönlichkeiten individueller Zielpersonen, sondern entsprechend der Konzeptualisierung als Wissenskonstrukt durchschnittliche Personen und typische Ergebnisse im Zusammenhang mit Persönlichkeitseigenschaften. Eine Überprüfung der auf Grund der drei Komponenten angenommenen dreifaktoriellen Struktur des Leistungstests wurde von Christiansen et al. nicht vorgenommen. Verwendet wurde auf Grund geringer Reliabilitäten der Subscores zudem nur ein DI-Gesamtwert, der sich in einem Pfadmodell als Prädiktor für akkurate

Persönlichkeitsbeurteilungen von fremden (β = .42) und bekannten (β = .41) Zielpersonen erwies (Christiansen et al., 2005).

De Kock et al. (2015) beschäftigten sich einige Jahre später erneut mit der DI. Sie übernahmen die Definition von Christiansen et al. (2005), bezeichneten das Konstrukt allerdings als Dispositional Reasoning, was in Anbetracht des Fokus auf Wissen keine passende Bezeichnung des Konstruktes darstellt. Aus diesem Grund wird im Folgenden weiterhin die Bezeichnung DI von Christiansen et al. verwendet. Neben der Benennung der drei Komponenten der DI (vgl. oben) sowie einer etwas umfassenderen Einordnung des Konstruktes in relevante Literatur aus der Persönlichkeitspsychologie überarbeiteten und ergänzten De Kock et al. den von Christiansen et al. konstruierten Leistungstest. Hierdurch wollten De Kock et al. eine Erfassung der drei angenommenen Komponenten ermöglichen und den Aufgabeninhalt an eine nicht-universitäre Stichprobe anpassen. In ihrer Studie konnten De Kock et al. zeigen, dass alle drei Komponenten einen Zusammenhang mit der Akkuratheit von Persönlichkeitsbeurteilungen aufweisen: $r = .14$ ($p = .11$) für Trait Induction, bis $r = .33$ ($p < .001$) für Trait Extrapolation sowie $r = .34$ ($p < .001$) für die gesamte DI. Trait Extrapolation sowie Trait Contextualization zeigten zudem inkrementelle Validität bei der Vorsage der Akkuratheit über die allgemeine Intelligenz hinaus (De Kock et al., 2015). Unter Verwendung einer anderen Methode zur Bestimmung der Akkuratheit (Differential Accuracy nach Cronbach, 1955; vgl. Abschnitt 2.3.1) zeigten sich allerdings andere Ergebnisse. Am auffälligsten war hierbei, dass die Trait Extrapolation unter Verwendung der zweiten Methode keinen Zusammenhang mehr zur Akkuratheit zeigte – der Zusammenhang der anderen beiden Komponenten sank hingegen in einem geringeren Ausmaß (De Kock et al., 2015). In einer späteren Veröffentlichung konnten De Kock et al. (2017) schließlich die angenommene dreifaktorielle Struktur des Messinstruments bestätigen, wobei die latenten Korrelationen mit .84 bis .95 sehr hoch ausfielen und ein einfaktorielles Modell ebenfalls einen akzeptablen, wenn auch signifikant schlechteren Fit zeigte (De Kock et al., 2017).

Auch de Vries et al. (2021) beschäftigtem sich mit der DI, allerdings unter dem wiederum neuen Namen Dispositional Insight (die Begründung für den neuen Namen – u.a. fehlende objektiv richtige Antworten – ist erneut nicht ganz nachvollziehbar). Sie entwickelten den Dispositional Insight Test (DIT), der nach Angaben der Autor:innen auf Inhalt und Format des Instruments von De Kock et al. (2015) aufbaut und sich von den Vorgängerinstrumenten insbesondere in dem Punkt unterscheidet, dass der DIT nicht auf dem FFM, sondern dem HEXACO-Modell basiert (de Vries et al., 2021). De Vries et al. präsentierten allerdings weder

Ergebnisse zur Dimensionalität des DIT, noch zu Zusammenhängen zur Akkuratheit von Persönlichkeitsbeurteilungen.

Insgesamt betrachtet ist bei dem Konstrukt DI eine gewisse Relevanz für akkurate Persönlichkeitsbeurteilungen offensichtlich. Zum einen liegt dies an dem Inhaltsbereich der DI, zum anderen an den vorhandenen empirischen Ergebnissen zum Zusammenhang mit der Akkuratheit von Persönlichkeitsbeurteilungen, die je nach Komponente der DI und Methode zur Bestimmung der Akkuratheit überwiegend klein bis mittel ausfielen (Christiansen et al., 2005; De Kock et al., 2015). Bei DI handelt es sich also um eine Fähigkeit, die bei akkuraten Persönlichkeitsbeurteilungen nicht unberücksichtigt bleiben sollte. Fraglich ist allerdings, ob DI als zentrale Fähigkeit betrachtet werden kann, die akkuraten Persönlichkeitsbeurteilungen von individuellen Zielpersonen zugrunde liegt. Hierbei müssen folgende Aspekte berücksichtigt werden: Die Komponente Trait Induction bezieht sich auf die Verbindung zwischen Persönlichkeitseigenschaft und Verhalten (Christiansen et al., 2005; De Kock et al., 2015). Als Beispiel nennen De Kock et al. (2015) das Wissen, dass eine gesprächige Person extravertiert ist und operationalisiert wird die Komponente durch eine Aufgabe, in der verschiedene Adjektive den Persönlichkeitsfaktoren zugeordnet werden müssen (Christiansen et al., 2005; De Kock et al., 2015; de Vries et al., 2021). Hier kann in Frage gestellt werden, inwieweit es für eine akkurate Beurteilung der Persönlichkeit einer individuellen Zielperson wichtig ist, Verhaltensweisen mit einem konkreten Label für eine Persönlichkeitseigenschaft in Verbindungen bringen zu können. Die Korrektheit des Labels unterscheidet sich je Persönlichkeitsmodell (z.B. FFM vs. HEXACO-Modell; Ashton & Lee, 2007; McCrae & John, 1992) und Persönlichkeitseigenschaften werden zudem in unterschiedlichen Modellen unterschiedlich konzeptualisiert (z.B. FFM und HEXACO-Modell vs. CAPS; Ashton & Lee, 2007; McCrae & John, 1992; Mischel & Shoda, 1995). Wie in Abschnitt 2.1 erläutert wurde, manifestieren sich Persönlichkeitseigenschaften im Verhalten der Zielperson, sodass die Beschreibung und Vorhersage relevanter Verhaltensweisen wichtiger erscheinen, als diese mit dem Label einer Persönlichkeitseigenschaft in Verbindung bringen zu können. Bei De Kock et al. (2015) ergaben sich für diese Komponente zudem bei allen durchgeführten Analysen die geringsten Zusammenhänge zur Akkuratheit von Persönlichkeitsbeurteilungen, was ebenfalls gegen eine Relevanz dieser Komponente spricht. Werden die Persönlichkeitsbeurteilungen hingegen eher abstrakt auf Ebene der Big Five-Faktoren oder deren Facetten erfasst, spielt die Fähigkeit der Trait Induction vermutlich eine deutlich wichtigere Rolle.

Auch die Komponente Trait Extrapolation (De Kock et al., 2015) kann kritisch betrachtet werden. Diese bezieht sich auf das Wissen, welche Persönlichkeitseigenschaften

normalerweise gemeinsam bei einer Person auftreten (Christiansen et al., 2005), was im Rahmen einer individuellen Persönlichkeitsbeurteilung nur dann zu einem korrekten Urteil führt, wenn die Zielperson dieser eher normativen Vorstellung entspricht. Wie De Kock et al. (2015) beschreiben, geht es bei dieser Komponente darum, fehlende Information über eine Persönlichkeitseigenschaft auf Basis der Information über andere Persönlichkeitseigenschaften zu ersetzen, was entsprechend mit einem Fehlurteil verbunden sein kann.

Die Komponente Trait Contextualization, die sich auf die Verbindung zwischen Persönlichkeitseigenschaften und Situationen bezieht (Christiansen et al., 2005; De Kock et al., 2015) hat für individuelle Persönlichkeitsbeurteilungen die vermutlich höchste Relevanz. Der Einfluss von Situationen auf das Verhalten wird auch in anderen Modellen thematisiert (Kammrath et al., 2005; Letzring & Funder, 2021) und muss von den Beurteiler:innen berücksichtigt werden, um zu entscheiden, ob ein Verhalten persönlichkeitsrelevant ist oder nicht. Allerdings muss berücksichtigt werden, dass es hierbei ebenfalls interindividuelle Unterschiede geben kann (vgl. Wenn-Dann-Profile; Kammrath et al., 2005; Mischel & Shoda, 1995). In der Studie von De Kock et al. (2015) zeigte die Trait Contextualization über alle durchgeführten Analysen hinweg die stabilsten Zusammenhänge zur Akkuratheit von Persönlichkeitsbeurteilungen, was deren mögliche Relevanz unterstützt.

2.2.6 Personale Intelligenz

In einer Veröffentlichung aus dem Jahr 2008 schlug Mayer das Konstrukt der *Personalen Intelligenz* (Personal Intelligence; PI) vor und definierte diese als „the capacity to reason about personality and to use personality and personal information to enhance one's thoughts, plans, and life experiences" (S. 210). Mayer nahm vier Fähigkeitsbereiche der PI an, die er wie folgt beschrieb (Mayer, 2008, 2009):

1. Persönlichkeitsrelevante Information erkennen
2. Auf Basis persönlichkeitsrelevanter Information akkurate Persönlichkeitsmodelle bilden
3. Persönlichkeitsrelevante Information nutzen, um Entscheidungen zu lenken
4. Ziele, Pläne und Lebensgeschichten systematisieren

Während anfänglich der Fokus der PI noch auf der eigenen Persönlichkeit zu liegen schien (vgl. Mayer, 2008), bezog sich der Autor wenig später gleichermaßen auf die eigene Persönlichkeit und die Persönlichkeit anderer Personen (Mayer, 2009). In einigen Veröffentlichungen wird die PI zudem kurz als Fähigkeit zum Schlussfolgern über

Persönlichkeit beschrieben und die vier Fähigkeitsbereiche als Bereiche des Problemlösens oder Schlussfolgerns (Mayer, 2015; Mayer et al., 2012, 2017, 2019).

Das Konzept PI baut auf verschiedenen Forschungsbereichen auf, wozu auch die bereits beschriebenen personalen Intelligenzen von Gardner (1983) gehören (Mayer, 2008, 2009). Mayer (2008) sah insbesondere Parallelen zu Gardners intrapersonalen Intelligenz und bezeichnete diese später als einen der Vorgänger der PI (Mayer, 2009). Auch der Moderator des Good Judge aus Funders (1995) Forschung zu akkuraten Persönlichkeitsbeurteilungen stellt Mayer (2008) zufolge einen für die PI relevanten Forschungsbereich dar. Dieser wurde von Mayer (2009) in den ersten Fähigkeitsbereich der PI einordnet, wobei eine Einordnung in den zweiten Bereich ebenso passend erscheint. Eine Abgrenzung der PI nahm Mayer (2008) wiederum zur SI sowie zur Emotionalen Intelligenz vor, wobei er letztere als möglichen Teil der PI und SI als Komplement der PI ansah (Mayer, 2009; vgl. auch Mayer et al., 2016).

Dass PI im Allgemeinen und insbesondere der zweite Fähigkeitsbereich für akkurate Persönlichkeitsbeurteilungen interessant erscheint, ist anhand der Beschreibungen von Mayer (2008, 2009) leicht zu erkennen. Allerdings muss beachtet werden, dass eine explizite und nachvollziehbare Herleitung der postulieren vier Fähigkeitsbereiche der PI bei Mayer (2008, 2009) nicht erkennbar ist und somit eine Erklärung, warum genau diese vier Bereiche angenommen werden, fehlt. Stattdessen verweist Mayer (2008) diesbezüglich lediglich auf übereinstimmende Aspekte verschiedener Persönlichkeitsmodelle (z.B. das Vorliegen stabiler Persönlichkeitseigenschaften), aus denen ihm zufolge die vier Fähigkeitsbereiche hervorgehen sollen (vgl. auch Mayer, 2009). Die vier Bereiche wiederum werden bei Mayer (2008, 2009) unter Verweis auf verschiedenste, größtenteils sehr spezifische Forschungsergebnisse und – wie der Autor selbst anmerkt – mit Hilfe weiterer noch spezifischerer Fähigkeiten beschrieben (u.a. Funder, 1995). Mayer (2008) zufolge werden durch sein Modell diverse Fähigkeiten organisiert und die entsprechende Forschung zu einem kohärenten Ganzen synthetisiert. Allerdings bleibt es überwiegend bei einer Aufzählung der spezifischen Forschungsergebnisse und Fähigkeiten, ohne dass eine Synthese deutlich wird. Aus diesem Grund liefert das Konzept PI nichts Neues im Vergleich zu den bereits in den vorherigen Kapiteln beschrieben Modellen und Theorien. Darüber hinaus bezieht sich der für akkurate Persönlichkeitsbeurteilungen interessanteste zweite Fähigkeitsbereich zwar unter anderem auch auf Persönlichkeitsmodelle anderer Personen, aus der Beschreibung von Mayer (2008) geht allerdings hervor, dass hier Modelle über durchschnittliche Persönlichkeiten gemeint sind und keine individuellen Personen. Hierfür spricht auch der später von Mayer (2015) vorgenommene Vergleich mit der

DI von Christiansen et al. (2005), die sich auf allgemeines Wissen über Persönlichkeit bezieht. Die Beschreibungen von Mayer aus dem Jahr 2009 sind wiederum uneindeutiger.

Mayer et al. (2012) entwickelten schließlich ein Testverfahren zur Erfassung von PI, den Test of Personal Intelligence (TOPI), der mittlerweile in der fünften Version vorliegt (Mayer et al., 2019). Für alle vier Fähigkeitsbereiche der PI wurden Multiple-Choice-Items entwickelt, von denen, wie bereits bei der DI, eine Antwortoption unter Verweis auf konkrete Forschungsergebnisse als korrekt gewertet wird (Mayer et al., 2012). Auch wenn nach eigenen Angaben der Autor:innen während der Entwicklung des TOPI darauf geachtet wurde, dass jeder Fähigkeitsbereich für sich genommen Eindimensionalität aufweist (Mayer et al., 2012), erwies sich eine empirische Trennung der vier angenommenen Bereiche der PI als schwierig (Mayer et al., 2012, 2017), sodass in der aktuellsten Version nur ein einziger dem Test zugrunde liegender Faktor angenommen wird (Mayer et al., 2019). Bei näherer Betrachtung der Items des TOPI (z.B. Mayer et al., 2012) fällt auf, dass einige Items Ähnlichkeit zu Items der Testverfahren zur Erfassung der DI aufweisen (vgl. Christiansen et al., 2005; de Vries et al., 2021). So werden sowohl die DI als auch der zweite Fähigkeitsbereich der PI mittels Items erfasst, in denen erfragt wird, welche Persönlichkeitseigenschaften üblicherweise gemeinsam bei einer Person auftreten (Christiansen et al., 2005; Mayer et al., 2012, 2019; vgl. auch Mayer, 2015). Und auch eine zweite Komponente der DI, das Wissen über die Verbindung von Persönlichkeitseigenschaft und Verhalten (Christiansen et al., 2005), findet sich in den Items zur Erfassung des dritten Fähigkeitsbereichs der PI wieder (vgl. Mayer et al., 2012, 2019). In Anbetracht dessen, dass PI im Kern Schlussfolgern oder Problemlösen darstellen soll (z.B. Mayer, 2015) und der Fokus von DI auf Wissen liegt (Christiansen et al., 2005), erscheint es nicht angemessen, dass beide Konstrukte mit sehr ähnlichen Items erfasst werden. Mayer et al. (2019) sprechen im Zusammenhang mit einer Itemart zur Erfassung des zweiten Fähigkeitsbereichs der PI sogar selbst von Trait Knowledge (siehe auch Mayer et al., 2012). Bei PI liegt somit eine gewisse Diskrepanz zwischen Konzeptualisierung und Operationalisierung vor, sofern Mayer (2008, Mayer et al., 2012) bei PI auch wirklich von Schlussfolgern im Sinne einer Fähigkeit mit Bezug zu Intelligenzstrukturmodellen (z.B. McGrew, 2009) ausgeht.

Zusammenfassend kann festgehalten werden, dass sich PI als theoretische Grundlage für eine Fähigkeit zur akkuraten Beurteilung der Persönlichkeit anderer Personen weniger eignet, da Unklarheiten bei der Konzeptualisierung vorhanden sind und PI keinen Mehrwert gegenüber bereits beschriebenen Theorien und Modelle (z.B. Funder, 1995) liefert. Ähnliches gilt für die vorhandene Operationalisierung von PI.

2.2.7 Akkurate Konstruktion und Verwendung von Wenn-Dann-Profilen

Die theoretischen Annahmen oder gewählten Operationalisierungen der bisher beschriebenen Forschung lassen erkennen, dass hier Persönlichkeit im Sinne des FFM konzeptualisiert und/oder als situationsübergreifend relativ konsistent angenommen wurde (vgl. Christiansen et al., 2005; Funder, 2012; Hall et al., 2017; Hirschmüller et al., 2013; Mayer, 2008; O'Sullivan et al., 1965; Süß et al., 2009). Akkurate Persönlichkeitsbeurteilungen werden allerdings ebenfalls im Zusammenhang mit dem in Abschnitt 2.1 beschriebenen CAPS von Mischel und Shoda (1995) und den dort angenommenen Wenn-Dann-Profilen der Persönlichkeit thematisiert. Für eine Beurteilung der Persönlichkeit müssen dem Ansatz zufolge zunächst Höhe und Form der Wenn-Dann-Profile einer Person identifiziert werden, um anschließend auf Basis dieser Information auf die zugrunde liegenden kognitiven und affektiven Einheiten (z.B. Ziele, Werte, Motive, Ansichten; Kammrath et al., 2005) und deren Interaktion zu schließen sowie eine Theorie über die Person zu erstellen (Mischel & Shoda, 1995). Kammrath et al. (2005) führten drei Studien durch, um zu untersuchen, ob Wenn-Dann-Profile für Persönlichkeitsbeurteilungen konstruiert und verwendet werden. In der ersten Studie wurden Zielpersonen mit Hilfe eines von fünf Persönlichkeitsadjektiven (u.a. freundlich, schüchtern) beschrieben. Anschließend sollte das Verhalten der Zielpersonen in verschiedenen interpersonalen Situationen vorhergesagt werden, wobei für alle Zielpersonen situationsspezifische Verhaltensvorhersagen in Form von Wenn-Dann-Profilen resultierten (Kammrath et al., 2005). Des Weiteren konnten die Autor:innen zeigen, dass Annahmen über die zugrunde liegende Motivation der Zielpersonen die Verhaltensvorhersage mediierte. In der zweiten und dritten Studie untersuchten Kammrath et al. die umgekehrte Assoziation. Hier wurden zunächst Wenn-Dann-Profile verschiedener Zielpersonen präsentiert und im Anschluss die Einschätzung bestimmter Persönlichkeitseigenschaften erhoben. Es ergaben sich in Abhängigkeit der präsentierten Wenn-Dann-Profile spezifische Einschätzungen von unter anderem Freundlichkeit und Schüchternheit sowie des Big Five-Faktors Verträglichkeit (nicht aber der Extraversion; Kammrath et al., 2005).

Kammrath et al. (2005) konnten somit aufzeigen, dass Wenn-Dann-Profile bei Persönlichkeitsbeurteilungen eine Rolle spielen – die Akkuratheit in der Konstruktion und Anwendung der Profile wurde von den Autor:innen allerdings nicht untersucht. Dies erfolgte nach eigenen Angaben erstmals durch Friesen und Kammrath (2011). Die Autorinnen konnten in ihrer Studie zeigen, dass Wenn-Dann-Profile mit einer gewissen Akkuratheit eingeschätzt werden können und hierbei interindividuelle Unterschiede vorliegen. Allerdings bezogen sich

die von Friesen und Kammrath untersuchten Wenn-Dann-Profile auf den sehr spezifischen Bereich der Auslösung von negativen Emotionen in engen Beziehungen und nicht auf den breiteren Bereich akkurater Persönlichkeitsbeurteilungen.

Auch wenn im Zusammenhang mit den Wenn-Dann-Profilen eine Fähigkeit bisher noch nicht thematisiert wurde, stellen das CAPS und die Wenn-Dann-Profile eine relevante theoretische Grundlage zur Einordnung und Konzeptualisierung einer Fähigkeit zur akkuraten Beurteilung der Persönlichkeit anderer Personen dar. Insbesondere durch die explizite Mitberücksichtigung situativer Aspekte bei der Konzeptualisierung von Persönlichkeit lassen sich die anderen Theorien und Modelle sinnvoll ergänzen. Im Hinblick auf eine Fähigkeit kann aus den obigen Studien (Friesen & Kammrath, 2011; Kammrath et al., 2005; Mischel & Shoda, 1995) zudem zusammenfassend abgeleitet werden, dass diesem Ansatz nach für akkurate Persönlichkeitsbeurteilungen die Wenn-Dann-Profile anderer Personen richtig konstruiert und angewendet werden müssen, beispielsweise um das Verhalten und Erleben von Personen vorherzusagen oder Persönlichkeitsbeurteilungen vorzunehmen.

2.2.8 Attributionstheorien

Gegenstand von Attributionstheorien sind die zugrundeliegenden Prozesse bei Schlussfolgerungen über die Ursachen von Verhalten (z.B. Gilbert et al., 1988; Trope, 1986; Trope & Liberman, 1993). In diesem Punkt wird insbesondere zwischen dispositionalen (z.B. Persönlichkeitseigenschaften, Motive, Fähigkeiten) und situativen Faktoren (z.B. soziale Norm, Aufgabenschwierigkeit) unterschieden (Gilbert & Malone, 1995; Trope, 1986), was die Relevanz für den Bereich akkurater Persönlichkeitsbeurteilungen deutlich macht. Im Folgenden sollen daher beispielhaft zwei Modelle sowie ein bekannter Beurteilungsfehler kurz beschrieben werden, die bereits in einigen der zuvor präsentierten Theorien und Modellen berücksichtigt wurden.

Trope (1986; vgl. auch Trope & Liberman, 1993) schlug ein Zwei-Phasen-Modell vor, in dem zwei Prozesse unterschieden werden, die jeder Schlussfolgerung über die Ursache von Verhalten zu Grunde liegen sollen: 1.) Identifikation und 2.) Dispositionale Schlussfolgerung. In der ersten Phase wird zunächst das Verhalten der Zielperson in Hinblick auf eine Disposition kategorisiert (z.B. das Verhalten der Zielperson ist verantwortungsbewusst), wobei neben dem Verhalten auch situative Faktoren und vorherige Information über die Zielperson (z.B. vergangenes Verhalten) berücksichtigt werden (Trope & Liberman, 1993). In der zweiten Phase des Modells werden in einem hypothesenprüfenden Prozess schließlich Rückschlüsse vom kategorisierten Verhalten auf die Dispositionen der Zielperson vorgenommen. Auch hier

erfolgt eine Anpassung an situative Einflüsse sowie eine Integration vorheriger Information (Trope & Liberman, 1993).

Das zweite Modell geht auf Gilbert et al. (1988) zurück und umfasst insgesamt drei Phasen des Attributionsprozesses: 1.) Kategorisierung des Verhaltens der Zielperson, 2.) Charakterisierung im Sinne eines Schlusses vom Verhalten der Zielperson auf die zugrunde liegende Disposition und 3.) Korrektur der Schlussfolgerung auf Basis situativer Information. Während die ersten beiden Prozesse relativ automatisch ablaufen sollen, wird der dritte Prozess als bewusster und relativ kontrollierter Prozess höherer Ordnung beziehungsweise eine Art des Schlussfolgerns beschrieben (Gilbert et al., 1988).

Ein im Hinblick auf akkurate Persönlichkeitsurteile interessantes Phänomen ist die sogenannte *Korrespondenzverzerrung* (Correspondence Bias; für eine Übersicht siehe Gilbert & Malone, 1995), die auch unter dem Namen *Fundamentaler Attributionsfehler* (Fundamental Attribution Error; Ross, 1977) bekannt ist. Die Korrespondenzverzerrung wird beschrieben als Tendenz, auch dann vom Verhalten einer Zielperson Rückschlüsse auf deren Disposition vorzunehmen, wenn das Verhalten vollständig durch situative Faktoren erklärt werden kann (Gilbert & Malone, 1995). Eine fast identische Beschreibung findet sich für den Fundamentalen Attributionsfehler, der die Tendenz darstellt, situative Einflüsse auf das Verhalten zu unterschätzen und dispositionale Einflüsse zu überschätzen (Ross, 1977).

Die Relevanz für akkurate Persönlichkeitsbeurteilungen ist anhand der Beschreibungen der beiden Attributionsmodelle und des Beurteilungsfehlers leicht ersichtlich. Allerdings thematisieren Attributionsmodelle oftmals nicht die Korrektheit der Attributionen und entsprechende interindividuelle Unterschiede (Funder, 1995). Ihre Bedeutung für akkurate Persönlichkeitsbeurteilungen ist daher vor allem in der Berücksichtigung situativer Faktoren als Ursache für Verhalten zu sehen und darin, dass nicht jedes Verhalten Rückschlüsse auf die Persönlichkeit der Zielperson erlaubt. Akkurate Persönlichkeitsbeurteilungen können den beschriebenen Modellen entsprechend nur dann resultieren, wenn die Beurteiler:innen auch situative Einflüsse auf das Verhalten berücksichtigen (z.B. indem Schlussfolgerungen auf die Persönlichkeitseigenschaft um einen situativen Einfluss korrigiert werden; Gilbert et al., 1988).

Dieser Aspekt wurde in einigen der zuvor beschrieben Modelle und Theorien bereits berücksichtigt. Die Untersuchung von Beurteilungsfehlern, wie dem Fundamentalen Attributionsfehler, ist Teil der Forschungsgeschichte zur Akkuratheit von Persönlichkeitsbeurteilungen, aus der heraus das RAM entstand (Funder, 1995, 1999). Letzring und Funder (2021) nennen Attributionstheorien zudem explizit als einen der Ausgangspunkte für die Entwicklung des RAM. Darüber hinaus spielt ihnen zufolge die Fähigkeit, situative

Ursachen für das Verhalten bei der Persönlichkeitsbeurteilung zu berücksichtigen, im letzten Schritt des RAM eine Rolle (Letzring & Funder, 2021; vgl. Abschnitt 2.2.2). Des Weiteren integriert das STAM als Erweiterung des RAM (vgl. Abschnitt 2.2.3) die beiden dargestellten Modelle von Trope (1986) und Gilbert et al. (1988) und nimmt eine Anpassung der Persönlichkeitsbeurteilung an kontextuelle Information explizit an (Hall et al., 2017). Auch die DI von Christiansen et al. (2005) basiert unter anderem auf dem Modell von Trope. Dieses wurde allerdings primär dafür verwendet, um die Bedeutung von Wissen über die Beziehung zwischen Verhalten und Persönlichkeitseigenschaften zu unterstreichen (Christiansen et al., 2005; De Kock et al., 2015). Für die Betonung der Wichtigkeit des situativen Einflusses auf Verhalten wurde später das Modell von Gilbert et al. und hier im Besonderen die dritte Phase (Korrektur) herangezogen (De Kock et al., 2015). Schließlich thematisieren Kammrath et al. (2005) im Zusammenhang mit den Wenn-Dann-Profilen Attributionsmodelle von unter anderem Trope und Gilbert et al. und betonen im Rahmen dessen vor allem einen zentralen Unterschied zwischen beiden Ansätzen: Während in den Attributionsmodellen angenommen wird, dass Persönlichkeit und Situation ausschließlich Haupteffekte auf das Verhalten ausüben und bei Persönlichkeitsbeurteilungen der situative Einfluss auf das Verhalten daher abgezogen werden muss, werden bei den Wenn-Dann-Profilen die situativen Einflüsse explizit mitberücksichtigt (Kammrath et al., 2005). Beide Ansätze betonen somit die Wichtigkeit des situativen Einflusses, wobei die Art und Weise, wie dieser berücksichtigt werden sollte, sich zwischen beiden Ansätzen maximal unterscheidet (Einfluss der Situation entfernen vs. mitberücksichtigen; vgl. Kammrath et al., 2005).

2.2.9 Weitere Modelle

Abschließend sollen kurz konzeptuelle Modelle und Theorien angesprochen werden, die für Persönlichkeitsbeurteilungen zwar prinzipiell große Relevanz besitzen und in der Literatur zum Teil stark verbreitet sind, die auf Grund der fehlenden Thematisierung interindividueller Unterschiede in der Akkuratheit oder einer Fähigkeit für das Ziel der vorliegenden Arbeit allerdings weniger relevant sind.

2.2.9.1 Weighted-Average Model. Eines dieser Modelle ist das Weighted-Average Model (WAM) von Kenny (1991; siehe auch Kenny, 1994). Das WAM nimmt neun Faktoren an, von denen die Übereinstimmung zwischen zwei Personen hinsichtlich der Beurteilung der (Persönlichkeits-)Eigenschaft einer Zielperson abhängt. Zu diesen Faktoren gehören laut Kenny (1991, 1994): 1.) Bekanntschaft zwischen Beurteiler:in und Zielperson, 2.)

Überlappung der behavioralen Information, die den Beurteiler:innen zur Verfügung steht, 3.) Konsistenz im Verhalten der Zielperson, 4.) geteilte Bedeutungssysteme der Beurteiler:innen bezüglich der Verhaltensweisen der Zielperson, 5.) Einfluss durch Stereotype des Erscheinungsbilds, 6.) Übereinstimmung der Stereotype zwischen den Beurteiler:innen, 7.) Validität der Stereotype, 8.) spezifischer Eindruck der beurteilenden Person von der Zielperson, 9.) Kommunikation zwischen den Beurteiler:innen. Laut Kenny (1994) handelt es sich beim WAM primär um ein theoretisches Modell, das den Prozess der Eindrucksbildung beschreiben soll. Es umfasst zudem ein mathematisches Modell, das die Bestimmung der Übereinstimmung zwischen zwei Beurteiler:innen erlaubt und stellt auch ein Maß für die Schätzung der Akkuratheit der Beurteilung zur Verfügung (Kenny, 1991, 1994). Wie Kenny (1994) betont, steht allerdings nicht im Fokus „who is accurate, but when and how people are accurate“ (S. 127), sodass interindividuelle Unterschiede in der Akkuratheit nicht thematisiert werden. Kenny (2004) nahm später eine Reparametrisierung des WAM in sechs Varianzquellen vor, die bei der Beurteilung der Persönlichkeit einer Zielperson durch eine beurteilende Person angenommen werden können. Das neue Modell, das Kenny (2004) PERSON nannte, und das WAM sind dem Autor zufolge formal identisch und erlauben identische Vorhersagen, sodass das PERSON-Modell nicht näher thematisiert wird.

2.2.9.2 Truth and Bias Model. Ein weiteres Modell, das zumindest kurz Erwähnung finden soll, ist das Truth and Bias Model (TBM) von West und Kenny (2011), das sowohl konzeptuelle Aspekte als auch einen Ansatz zur Bestimmung und Analyse akkurater Persönlichkeitsbeurteilungen umfasst. Das TBM ist nicht auf Persönlichkeitsbeurteilungen beschränkt, sondern soll als integratives Rahmenkonzept die Untersuchung von Akkuratheit und Bias in der menschlichen Wahrnehmung in verschiedensten Bereichen ermöglichen (West & Kenny, 2011). Die Grundannahme des TBM besteht nach West und Kenny darin, dass Urteile sowohl von der Wahrheit (z.B. tatsächliche Ausprägung einer bestimmten Persönlichkeitseigenschaft) als auch von weiteren systematischen Faktoren (Bias; z.B. Ausprägung der einzuschätzenden Persönlichkeitseigenschaft bei der beurteilenden Person) beeinflusst werden und die Stärke dieser beiden Einflüsse variieren kann. Wahrheit und Bias können das Urteil hierbei in die gleiche Richtung oder in gegensätzliche Richtungen lenken, sodass ein Bias sowohl zu akkurateren als auch inakkurateren Urteilen führen kann (West & Kenny, 2011). Des Weiteren werden im TBM Moderatorvariablen angenommen, die die Einflussstärke von Wahrheit und Bias verändern können. Als Beispiel für eine Moderatorvariable, die die Einflussstärke der Wahrheit erhöhen und gegebenenfalls zusätzlich

die Einflussstärke des Bias reduzieren kann, nennen West und Kenny unter anderem Funders (1995) Moderator des Good Judge, der – wie in Abschnitt 2.2.2 dargestellt – eine Fähigkeitskomponente umfasst. Das TBM liefert somit einen Rahmen zur Einordnung, auf welche Art und Weise eine Fähigkeit die Akkuratheit von Persönlichkeitsbeurteilungen beeinflussen kann, allerdings ohne näher auf die zugrundeliegenden kognitiven Prozesse einzugehen. Darauf aufbauend wurde von West und Kenny ein statistischer Ansatz unter Verwendung multipler Regressionsanalysen vorgeschlagen. Dieser soll die Untersuchung des Einflusses von Wahrheit, Bias und Moderatorvariablen auf die Urteile sowie die Bestimmung und Analyse der Akkuratheit von Persönlichkeitsbeurteilungen ermöglichen (West & Kenny, 2011).

2.2.10 Zusammenfassung und Fazit

Insgesamt betrachtet sind von den in Abschnitt 2.2 vorgestellten konzeptuellen Modellen und Theorien folgende für die Einordnung und Konzeptualisierung einer Fähigkeit zur akkuraten Beurteilung der Persönlichkeit anderer Personen von Interesse: Das RAM (Funder, 1995), SI (insb. Weis & Süß, 2005), DI (Christiansen et al., 2005), Attributionsmodelle (Gilbert et al., 1988; Trope, 1986) sowie das CAPS (Kammrath et al., 2005; Mischel & Shoda, 1995).

Dem RAM von Funder (1995) kann entnommen werden, dass die Fähigkeit in den letzten Schritt des Persönlichkeitsbeurteilungsprozesses eingeordnet werden kann, in dem bereits Hinweise auf die Persönlichkeitseigenschaft vorliegen, die korrekt genutzt werden müssen. Welche Teilfähigkeiten hier eine Rolle spielen könnten, wird zudem von Letzring und Funder (2021) beschrieben. Die Autorin und der Autor nennen hierbei auch die Notwendigkeit zur Berücksichtigung situativer Einflüsse auf das Verhalten, was zusätzlich durch die Attributionsmodelle von Gilbert et al. (1988) und Trope (1986) unterstützt wird. Dem Integrativen Modell Sozialer Intelligenz von Weis und Süß (2005; vgl. auch Weis, 2008) zufolge können korrekte Schlussfolgerungen über Persönlichkeitseigenschaften als Teilleistung von SU angesehen werden und auch die Arbeiten von Vernon (1933), Wedeck (1947) und O'Sullivan und Guilford (1975; O'Sullivan et al., 1965) machen deutlich, dass akkurate Persönlichkeitsbeurteilungen als Teilbereich der SI angesehen werden können. Insbesondere die Fähigkeit SU liefert einen Hinweis darauf, dass die Fähigkeit zum logischen Schlussfolgern bei Persönlichkeitsbeurteilungen eine wichtige Rolle spielen könnte (vgl. Weis, 2008). Offen ist allerdings die Frage, inwieweit eine inhaltliche Abgrenzung von SU und einer Fähigkeit zur akkuraten Persönlichkeitsbeurteilung möglich ist. Das CAPS und die dort beschriebenen Wenn-Dann-Profile (Kammrath et al., 2005; Mischel & Shoda, 1995) ergänzen

den theoretischen Hintergrund insofern, dass situative Aspekte nicht nur als alternative Einflüsse auf das Verhalten einer Zielperson berücksichtigt werden sollten, sondern auch in der Konzeptualisierung von Persönlichkeit und somit auch in deren Beurteilung miteinbezogen werden können. Auch die DI von Christiansen et al. (2005) hat für akkurate Persönlichkeitsbeurteilungen Relevanz. Da der Fokus hier allerdings auf allgemeinem Wissen über Persönlichkeit und nicht individuellem Wissen über einzelne Zielpersonen liegt, ist die Frage, inwieweit DI wirklich einen zentralen Prozess bei der Persönlichkeitsbeurteilung widerspiegelt. Die Diskussion der offenen Fragen sowie eine Integration der als relevant erachteten Modelle und Theorien im Hinblick auf die Frage, wie die zentrale Fähigkeit im Prozess akkurater Persönlichkeitsbeurteilungen beschrieben werden kann, wird in Kapitel 3 (Abschnitt 3.1) vorgenommen.

Die interpersonale Intelligenz von Gardner (1991) sowie die PI von Mayer (2008, 2009) stellen ebenfalls Theorien dar, die im Hinblick auf eine Fähigkeit zur akkuraten Persönlichkeitsbeurteilung auf den ersten Blick nützlich erscheinen. Allerdings weisen beide Theorien nur eine unzureichende theoretische Fundierung, Unklarheiten bei der Konzeptualisierung sowie keinen Mehrwert auf (vgl. Abschnitte 2.2.4.3 und 2.2.6). In der ursprünglichen Version des Linsenmodells von Brunswik (1956) wird hingegen keine Fähigkeit berücksichtigt. Da es sich beim RAM aber um eine Erweiterung des Linsenmodells handelt (Funder, 1999), erscheint eine zusätzliche Berücksichtigung auch nicht erforderlich. Beim STAM (Hall et al., 2017) fehlt es vor allem an empirischer Evidenz und das WAM (Kenny, 1991), das PERSON-Modell (Kenny, 2004) sowie das TBM (West & Kenny, 2011) diskutieren ebenfalls keine Fähigkeit und sind daher nicht weiter von Interesse.

2.3 Modelle und Methoden zur Bestimmung und Analyse

Nach den primär konzeptuellen Modellen und Theorien sollen nun Modelle und Methoden angesprochen werden, die die Bestimmung und Analyse der Akkuratheit von Persönlichkeitsbeurteilungen fokussieren. Wie zu sehen sein wird, besteht keine Einigkeit in der Frage, wie genau die Akkuratheit bestimmt und analysiert werden sollte. Zudem werden die bestimmten Akkuratheitswerte oftmals als Operationalisierungen einer theoretisch nicht näher beschriebenen Fähigkeit angesehen (z.B. Biesanz, 2010; Cronbach, 1955; Funder, 1999; Hall et al., 2018; siehe aber auch Christiansen et al., 2005). Gemeinsam haben allerdings alle Modelle und Methoden, dass zur Bestimmung des Ausmaßes der Akkuratheit die wahre Persönlichkeitsausprägung der Zielperson benötigt wird, was auf Grund des eigentlich latenten

Merkmales eine große Herausforderung darstellt (Funder, 1999). Vorgeschlagen wurden bisher folgende Kriterien, die als Indikatoren für die latente Persönlichkeitsausprägung dienen und somit auch die Bestimmung der Akkuratheit ermöglichen sollen:

1. Selbstbericht: Das am häufigsten verwendete Kriterium ist die Selbstauskunft der Zielperson über ihre eigene Persönlichkeit, beispielsweise erfasst durch ein standardisiertes Persönlichkeitsinventar (Kenny, 1994; Funder, 2012; vgl. auch Conzelmann et al., 2013). Der resultierende Akkuratheitswert wird in der Regel als Selbst-Fremd Übereinstimmung bezeichnet (z.B. Funder, 2012).
2. Konsensus bzw. Fremd-Fremd Übereinstimmung: Konsensus beschreibt die Übereinstimmung von mindestens zwei Beurteiler:innen hinsichtlich der Persönlichkeitsbeurteilung derselben Zielperson (Funder, 2012). Um das Ausmaß der Akkuratheit einer einzelnen Persönlichkeitsbeurteilung zu bestimmen, kann als Kriterium die mittlere Einschätzung einer Gruppe von Beurteiler:innen verwendet werden (Kenny, 1994; vgl. auch Legree et al., 2005).
3. Expert:innenurteil: Als Kriterium wird die Persönlichkeitsbeurteilung einer Person verwendet, von der man ausgehen kann, dass sie die Zielperson und ihre wahre Persönlichkeit sehr gut kennt (z.B. Eltern, Partner:in, enge Freund:innen; Kenny, 1994). Es besteht zudem die Möglichkeit, die mittlere Einschätzung mehrerer solcher Expert:innen zu verwenden (Legree et al., 2005).
4. Behaviorale Kriterien: Als Kriterium werden Verhaltensweisen oder verhaltensbezogene Lebensereignisse der Zielperson verwendet, die für die einzuschätzende Persönlichkeitseigenschaft relevant sind (Bernieri, 2001; Funder, 1999, 2012; Kenny, 1994). Die Verwendung solcher behavioraler Kriterien kann auf zwei Arten erfolgen: Einerseits kann betrachtet werden, inwieweit Persönlichkeitsbeurteilungen diese Verhaltenskriterien vorhersagen können (Funder, 2012), andererseits wurde vorgeschlagen, die Verhaltensweisen direkt und anstelle der Persönlichkeitseigenschaften von den Beurteiler:innen einschätzen zu lassen (Kenny, 1994).
5. Operationale Kriterien: Hierbei handelt es sich um per Definition (z.B. zwei Personen sind ein Paar) oder experimenteller Manipulation (z.B. Person wurde instruiert zu lügen) festgelegte Kriterien (Bernieri, 2001; Kenny, 1994). Diese erlauben somit eine objektive Bewertung der Beurteilungen als richtig oder falsch. Allerdings sind operationale Kriterien im Bereich der Persönlichkeitsbeurteilungen

laut Kenny (1994) kaum anwendbar und spielen bei den im Folgenden vorgestellten Modelle und Methoden daher keine Rolle.

An dieser Stelle sei schon einmal erwähnt, dass alle zuvor beschriebenen Kriterien problembehaftet sind (Funder, 1999, 2012; Kenny, 1994), was in Abschnitt 3.2 noch genauer erläutert wird. Daher wurde von Funder (1995) eine Kombination mehrerer dieser Kriterien empfohlen. Funder (1995; vgl. auch Funder & West, 1993) verglich die Evaluation der Akkuratheit in dem Zusammenhang mit dem Vorgang der Konstruktvalidierung hinsichtlich der beurteilten Persönlichkeitseigenschaft.

Die im Folgenden beschriebenen Modelle und Methoden basieren überwiegend auf Korrelations- und Regressionsanalysen. Der Fokus liegt dabei auf Ansätzen, die die Erfassung von Unterschieden in der Akkuratheit zwischen Beurteiler:innen erlauben. Ansätze, bei denen die Persönlichkeitsbeurteilungen über mehrere Beurteiler:innen hinweg gemittelt werden (vgl. Funder, 1999; Kenny & Winquist, 2001; Nestler & Back, 2017), werden – mit einer Ausnahme – nicht näher thematisiert.

Eine weitere Unterscheidung, die hinsichtlich der Modelle und Methoden gemacht werden muss, betrifft die Frage, ob die Akkuratheit je beurteilter Zielperson (*Profile Accuracy*) oder je beurteilter Persönlichkeitseigenschaft (*Trait Accuracy*) bestimmt wird (Back & Nestler, 2016; Connelly & Ones, 2010; Hall et al., 2018). Bei Verwendung einfacher Korrelationsanalysen werden zur Bestimmung der Profile Accuracy die Beurteilungen aller Persönlichkeitseigenschaften einer Zielperson mit den jeweiligen Kriteriumswerten korreliert, sodass ein Wert für jede Kombination aus Beurteiler:in und Zielperson resultiert (Back & Nestler, 2016; Connelly & Ones, 2010; Funder, 1999), der auch als Profilkorrelation bezeichnet wird (z.B. Back & Nestler, 2016). Im Kontrast dazu werden bei der Trait Accuracy die Beurteilungen aller Zielpersonen hinsichtlich einer Persönlichkeitseigenschaft mit den jeweiligen Kriteriumswerten korreliert, sodass eine Korrelation je Beurteiler:in und Persönlichkeitseigenschaft resultiert (Back & Nestler, 2016; Connelly & Ones, 2010). Laut Back und Nestler (2016) werden beide Ansätze oftmals als austauschbare Varianten verwendet. Den Autoren zufolge bestehen zwischen beiden Ansätzen aber wichtige konzeptuelle Unterschiede, sodass sie als unterschiedliche psychologische Phänomene aufgefasst werden sollten. Hierauf aufbauend nehmen Hall et al. (2018) an, dass der Profile Accuracy und Trait Accuracy zwei konzeptuell verschiedene Vergleichsprozesse zugrunde liegen, die verschiedene Fähigkeiten widerspiegeln. So sei bei der Profile Accuracy die Fähigkeit involviert, bei einzelnen Zielpersonen zwischen den relativen Ausprägungen verschiedener Persönlichkeitseigenschaften zu unterschieden und bei der Trait Accuracy die Fähigkeit,

verschiedene Zielpersonen im Hinblick auf deren Ausprägungen bei einer einzelnen Persönlichkeitseigenschaft zu unterscheiden (Hall et al., 2018; vgl. auch Back & Nestler, 2016). Eine weitere theoretische Einordnung der angenommenen Fähigkeiten wurde nicht vorgenommen. Hall et al. (2018) spekulieren allerdings über das Vorliegen verschiedener kognitiver Anforderungen sowie über möglicherweise involvierte kognitive Stile. Zudem ergaben insgesamt fünf von den Autor:innen durchgeführte Studien, dass beide Ansätze im mittel nur zu $\bar{r}$ = .30 (p < .001) miteinander korrelieren (Hall et al., 2018). Darüber hinaus zeigte sich über die fünf Studien hinweg, dass die unter Verwendung des Trait Accuracy-Ansatzes bestimmten Akkuratheitswerte der Big Five-Faktoren untereinander kaum substantielle Korrelationen über .10 aufweisen. Aus diesem Grund nahmen Hall et al. an, dass die bei der der Trait Accuracy zugrunde liegende, angenommene Fähigkeit ebenfalls weiter differenziert werden sollte.

In Abschnitt 2.2 wurden konzeptuelle Modelle und Theorien vorgestellt, die eine Erfassung interindividueller Unterschiede in der Akkuratheit von Persönlichkeitsbeurteilungen ermöglichen. Die Analysen, die mit Bezug auf Brunswiks (1956) Linsenmodell vorgeschlagen wurden (Nestler & Back, 2017; Tucker, 1964), sind dem Ansatz der Trait Accuracy zuzuordnen (Back & Nestler, 2016; Nestler & Back, 2017). Gleiches gilt für die Analysen des TBM von West und Kenny (2011). Wie berichtet wurde, sind globale Persönlichkeitsbeurteilungen auch Teil der Operationalisierung des Integrativen Modells Sozialer Intelligenz von Weis und Süß (2005). Im MTSI werden unter anderem die Big Five-Faktoren beurteilt und die Auswertung erfolgt je Zielperson über alle Big Five hinweg unter Verwendung von Distanzwerten (Süß et al., 2009), sodass hier der Ansatz der Profile Accuracy verwendet wurde.

2.3.1 Cronbach's Components of the Accuracy Score

Auf die zuletzt beschriebene Bestimmung der Profile Accuracy mit Hilfe sogenannter Distanzwerte bezieht sich eine vielfach in der Literatur zitierte Kritik von Cronbach aus dem Jahr 1955, die nach Ansicht einiger Autor:innen (z.B. Biesanz, 2010; Funder & West, 1993; Kenny, 1994) mitverantwortlich für eine abrupte Pause in der Forschung zu akkuraten Persönlichkeitsbeurteilungen war.

Distanzwerte wurden und werden oftmals bestimmt, indem je Beurteiler:in die quadrierten Distanzen zwischen den Fremdbeurteilungen der Zielperson auf einzelnen Persönlichkeitsitems sowie den entsprechenden Kriteriumswerten berechnet und im Anschluss alle quadrierten Distanzen über alle Items und Zielpersonen hinweg gemittelt werden (Cronbach, 1955; vgl. auch Legree et al., 2010; Legree et al., 2005). Cronbach (1955) konnte

zeigen, dass sich ein solcher globaler Akkuratheitswert aus den folgenden vier Komponenten zusammensetzt:

1. Die Komponente *Elevation* repräsentiert nach Cronbach (1955) Unterschiede in der Nutzung der Rating-Skala zwischen Beurteiler:in und allen Zielpersonen, das heißt Unterschiede zwischen dem Mittelwert aller Beurteilungen und dem Mittelwert aller Kriteriumswerte (über alle Zielpersonen und Persönlichkeitsitems hinweg).
2. Die Komponente *Differential Elevation* repräsentiert nach Cronbach (1955) die Fähigkeit, Abweichungen des durchschnittlichen Selbstberichts (d.h. Mittelwert über alle Persönlichkeitsitems hinweg) der einzelnen Zielpersonen vom Durchschnitt aller Zielpersonen korrekt zu beurteilen.
3. Die Komponente *Stereotype Accuracy* spiegelt nach Cronbach (1955) die Fähigkeit wider, Form und Streuung des durchschnittlichen Persönlichkeitsprofils aller Zielpersonen korrekt zu beurteilen.
4. Die Komponente *Differential Accuracy* steht schließlich für die Fähigkeit, Unterschiede zwischen den Zielpersonen in allen beurteilten Persönlichkeitsitems korrekt abzubilden (Cronbach, 1955). Nach Cronbach (1955) entspricht dieser Wert der Übereinstimmung zwischen allen Beurteilungen und Kriteriumswerten, nachdem die vorherigen drei Komponenten kontrolliert wurden.

Laut Cronbach (1955) spiegeln nur die Differential Elevation sowie die Differential Accuracy – beziehungsweise eine jeweils enthaltene korrelative Komponente – Fähigkeiten wider, die im Rahmen akkurater Persönlichkeitsbeurteilungen näher betrachtet werden sollten. Im Gegensatz dazu sollte die Elevation-Komponente in der Regel vollständig eliminiert (siehe auch Cronbach, 1958) und die Stereotype Accuracy näher untersucht werden.

Auch wenn Cronbach (1955) sich in seiner Kritik primär auf Distanzwerte bezog, betrifft diese zum Teil auch die weiter oben beschriebenen Profilkorrelationen, die eine spezielle Variante der Distanzwerte darstellen (Cronbach & Gleser, 1953; Legree et al., 2010). Während die Elevation-Komponente hier keine Rolle spielt (Cronbach & Gleser, 1953; Legree et al., 2010), beinhalten die Profilkorrelationen weiterhin die Stereotype Accuracy-Komponente (Funder, 1999; Furr, 2008; Kenny & Winquist, 2001). Als eine Möglichkeit zur Kontrolle der Stereotype Accuracy-Komponente schlugen Kenny und Winquist (2001; vgl. Furr, 2008 für einen ähnlichen Ansatz) das Zentrieren der einzelnen Beurteilungen und Kriteriumswerte anhand der jeweiligen Mittelwerte vor. Nach Funder (1999; vgl. auch Vogt & Colvin, 2003) eignen sich auch semipartielle Korrelationen, bei denen das durchschnittliche Persönlichkeitsprofil der Zielpersonen auspartialisiert wird. Funder erläutert allerdings auch

potentielle Nachteile der Nutzung von Semipartialkorrelationen. So kann es ihm zufolge vorkommen, dass durch das Auspartialisieren neben Fehlervarianz auch wahre Varianz entfernt wird und folglich verzerrte Akkuratheitswerte entstehen. Das kann insbesondere dann passieren, wenn Zielpersonen beurteilt werden, die ein durchschnittliches Persönlichkeitsprofil aufweisen (Colvin & Bundick, 2001; Funder, 1999). Zudem kann die Stereotype Accuracy laut Vogt und Colvin (2003) auch einen validen Prozess der Persönlichkeitsbeurteilung darstellen und in bestimmten Situationen die vorliegenden behavioralen Hinweise sinnvoll ergänzen. Ergebnisse ihrer Studie deuten den Autoren zufolge darauf hin, dass die akkuratesten Beurteiler:innen in der Studie nur dann Wissen über das durchschnittliche Persönlichkeitsprofil angewendet haben, wenn die Zielperson auch ein solches durchschnittliches Profil aufwies. Beispielsweise ergab sich nur dann ein Zusammenhang zwischen der Akkuratheit der Persönlichkeitsbeurteilungen – ermittelt durch Profilkorrelationen – und der Stereotype Accuracy, wenn die Zielperson auch ein durchschnittliches Persönlichkeitsprofil aufwies (Vogt & Colvin, 2003). Ob und wie eine Kontrolle der Stereotype Accuracy bei Profilkorrelationen vorgenommen wird, sollte nach Funder (1999) letztendlich in Abhängigkeit von der konkreten Forschungsfrage entschieden werden. Ein Beispiel hierfür findet sich bei Letzring (2008).

Einige Jahre später distanzierte sich Cronbach (1958) vollständig von seinem Ansatz (vgl. Cronbach, 1992) und sprach sich unter anderem dagegen aus, einen einzigen globalen Akkuratheitswert über alle Persönlichkeitseigenschaften hinweg zu bestimmen. Sein Komponentenmodell und insbesondere die Stereotype Accuracy-Komponente hatte allerdings trotz seiner späteren Distanzierung großen Einfluss auf weitere Modelle, wie im Folgenden zu sehen sein wird.

2.3.2 Social Relations Model

Einen anderen Weg, mit der Kritik von Cronbach (1955) umzugehen, wählte Kenny mit dem Social Relations Model (SRM; Kenny, 1994). So werden im SRM Komponenten angenommen, die Kenny (1994) zufolge denen von Cronbach ähneln. Zudem wurde von Kenny der nomothetische Fokus seines Ansatzes, in dem interindividuelle Unterschiede nicht von Interesse sind, betont. Aus diesem Grund ist das SRM für das Ziel der vorliegenden Arbeit weniger relevant. Da es allerdings im Social Accuracy Model von Biesanz (2010), das im folgenden Kapitel näher beschrieben wird, eine wichtige Rolle spielt, werden die hierfür zentralen Aspekte des SRM im Folgenden kurz erläutert.

Das SRM ist laut Kenny (1994) ein allgemeines Modell der Personenwahrnehmung und beschäftigt sich damit, wann und wie akkurate Einschätzungen zustande kommen. Dem

Modell zufolge setzt sich die Wahrnehmung einer bestimmten Eigenschaft einer anderen Person aus vier Komponenten beziehungsweise Varianzquellen zusammen (Kenny, 1994):

1. *Constant*: Die mittlere Wahrnehmung der Eigenschaft über alle Zielpersonen und Beurteiler:innen hinweg.
2. *Perceiver Effect*: Diese Komponente beschreibt, inwieweit ein:e Beurteiler:in dazu tendiert, die Eigenschaft bei allen Zielpersonen eher als hoch oder niedrig ausgeprägt wahrzunehmen.
3. *Target Effect*: Diese Komponente beschreibt, inwieweit die Eigenschaft bei einer Zielperson durch alle Beurteiler:innen eher als hoch oder niedrig ausgeprägt wahrgenommen wird.
4. Der *Relationship Effect* beschreibt die individuelle Wahrnehmung der Zielperson durch eine:n Beurteiler:in. Dieser Effekt ist bei nur einem Messzeitpunkt nicht von einer Messfehlerkomponente zu trennen.

Um die Varianzkomponenten untersuchen und quantifizieren zu können, wird ein Untersuchungsdesign benötigt, in dem mehrere Personen sich gegenseitig hinsichtlich einer bestimmten Eigenschaft beurteilen (Round-Robin-Design; Kenny, 1994). Akkuratheit wird im SRM nicht für einzelne Beurteiler:innen bestimmt, sondern für einzelne Eigenschaften und über alle Beurteiler:innen (und Zielpersonen) hinweg (Kenny, 1994). Hier werden laut Kenny (1994) analog zu Cronbach (1955) vier Typen der Akkuratheit unterschieden, die sich auf die vier beschriebenen Komponenten beziehen: Elevation Accuracy, Perceiver Accuracy, Generalized Accuracy, Dyadic Accuracy.

2.3.3 Social Accuracy Model

Beim Social Accuracy Model (SAM) von Biesanz (2010) handelt es sich nach Angaben des Autors um eine Integration des Komponentenansatzes von Cronbach (1955) und des SRM von Kenny (1994). Diese Integration resultierte in einem relativ umfassendem Komponentenmodell zur Analyse interindividueller Unterschiede in der Akkuratheit von Persönlichkeitsbeurteilungen, das berücksichtigt, dass es sowohl unterschiedlich gute Beurteiler:innen als auch unterschiedlich gute Zielpersonen gibt (Biesanz, 2010). Die Beurteilung der Akkuratheit erfolgt laut Biesanz zudem in Anlehnung an Funders (1995) Vergleich der Persönlichkeitsbeurteilung mit dem Vorgang der Konstruktvalidierung hinsichtlich der beurteilten Persönlichkeitseigenschaft, der für die Verwendung multipler Kriterien spricht.

Ausgangspunkt des SAM ist laut Biesanz (2010) das Urteil einer Person hinsichtlich verschiedener Persönlichkeitseigenschaften einer Zielperson, dessen Akkuratheit – die sogenannte *Impressionistic Accuracy* – anhand eines oder mehrerer Kriterien beurteilt werden soll. Das SAM ist somit dem Ansatz der Profile Accuracy zuzuordnen (Back & Nestler, 2016). Die Impressionistic Accuracy ist Biesanz zufolge allerdings nicht der primäre Fokus des Modells, sondern wird zum einen zerlegt in einen Effekt der beurteilenden Person (*Perceptive Accuracy*) und einen Effekt der Zielperson (*Expressive Accuracy*). Bei der Perceptive Accuracy handelt es sich um „the extent to which a particular perceiver's impressions are more or less accurate than other perceivers on average across different targets" (S. 861) und bei der Expressive Accuracy um „the extent to which a particular target is accurately perceived on average across different perceivers" (Biesanz, 2010, S. 861). Zum anderen erfolgt im SAM eine Zerlegung in die Stereotype Accuracy und Differential Accuracy nach Cronbach (1955), die im SAM als *Normative Accuracy* und *Distinctive Accuracy* bezeichnet und sowohl für Beurteiler:in als auch Zielperson modelliert werden (Biesanz, 2010). Die normativen Komponenten (*Perceiver Normative Accuracy*, *Target Normative Accuracy*) beziehen sich laut Biesanz auf die durchschnittlichen Ausprägungen der beurteilten Persönlichkeitseigenschaften und die distinktiven Komponenten (*Perceiver Distinctive Accuracy*, *Target Distinctive Accuracy*) auf die hiervon abweichenden individuellen Eigenschaften der beurteilten Zielperson. Im SAM wird das Urteil über mehrere Persönlichkeitseigenschaften hinweg in einem Regressionsmodell als abhängige Variable und das Kriterium oder die Kriterien zur Beurteilung der Akkuratheit als unabhängige Variable modelliert. Um die normativen von den distinktiven Komponenten der Akkuratheit zu trennen, wird zudem die durchschnittliche Ausprägung der beurteilten Persönlichkeitseigenschaften als weitere unabhängige Variable in das Modell mit aufgenommen. Die Schätzung individueller Unterschiede der beiden Komponenten für Beurteiler:innen und Zielpersonen erfolgt schließlich durch Aufstellung eines Mehrebenenmodells (Biesanz, 2010). Darüber hinaus ermöglicht das SAM laut Biesanz die Untersuchung möglicher Moderatoren der Akkuratheitskomponenten.

Von besonderem Interesse hinsichtlich einer für akkurate Persönlichkeitsbeurteilungen relevanten Fähigkeit ist die distinktive Komponente der Perceptive Accuracy, die von Biesanz als „extent to which one perceives the distinct, unique characteristics of others" definiert wurde (Biesanz, 2010, S. 864). Nach Biesanz (2010) spiegelt diese Komponente die klassischen guten Beurteiler:innen der Persönlichkeit wider. Zudem setzte Biesanz diese Komponente mit einer Fähigkeit gleich, die individuellen Persönlichkeitseigenschaften anderer zu entdecken – allerdings ohne eine weitere theoretische Einordnung vorzunehmen. Nach Human und Biesanz

(2011) ist die Perceiver Distinctive Accuracy zudem gleichzusetzen mit einer akkuraten Persönlichkeitsbeurteilung nach dem RAM von Funder (1995) und erfordert individuelles Wissen über die zu beurteilende Person (vs. allgemeines Wissen über Persönlichkeit bei der Normative Accuracy; vgl. auch Biesanz, 2010).

In zwei Studien fand Biesanz (2010) für die Perceiver Distinctive Accuracy nur relativ geringe interindividuelle Unterschiede, während sich für die anderen drei Komponenten (Perceiver Normative Accuracy, Target Distinctive Accuracy, Target Normative Accuracy) entsprechende Unterschiede zeigten. Die Ergebnisse deuten dem Autor zufolge darauf hin, dass vor allem eine Heterogenität in der Akkuratheit, mit der verschiedene Zielpersonen beurteilt werden, vorliegt, woraus Biesanz die Notwendigkeit weiterer Forschung mit Fokus auf die Zielpersonen ableitete. Unter Bezugnahme auf dieses Ergebnis sowie die ersten beiden Schritte des RAM von Funder (1995) untersuchten Rogers und Biesanz (2019) die Rolle guter Zielpersonen (d.h. der Target Distinctive Accuracy) bei der Erfassung interindividueller Unterschiede in der Perceiver Distinctive Accuracy. Die Autorin und der Autor fanden in insgesamt vier Studien Interaktionseffekte zwischen beiden Komponenten, die darauf hindeuten, dass gute Beurteiler:innen insbesondere bei guten Zielpersonen akkuratere Urteile erzielen. Rogers und Biesanz schlussfolgerten auf Basis ihrer Ergebnisse, dass interindividuelle Unterschiede in einer Fähigkeit zu akkuraten Persönlichkeitsurteilen existieren, sich allerdings nur dann zeigen, wenn gute Zielpersonen beurteilt werden. Im Kontrast dazu werde eine schlechte Zielperson von guten Beurteiler:innen nicht akkurater eingeschätzt als von schlechten Beurteiler:innen (Rogers & Biesanz, 2019). Rogers und Biesanz lieferten mit diesem Ergebnis zudem Evidenz für das RAM (Funder, 1995), das besagt, dass bei fehlenden persönlichkeitsrelevanten Hinweisen (d.h. bei schlechten Zielpersonen) keine akkurate Persönlichkeitsbeurteilung möglich ist. Darüber hinaus zeigten sich die Ergebnisse der Autor:innen sowohl in zwei der insgesamt vier durchgeführten Studien, in denen Beurteiler:innen und Zielpersonen direkt miteinander interagieren konnten, als auch in den anderen beiden Studien, in denen Videos der Zielpersonen präsentiert und die verfügbaren Hinweise somit konstant gehalten wurden (Rogers & Biesanz, 2019). Daher schlussfolgerten Rogers und Biesanz, dass die Leistung der guten Beurteiler:innen auf ein besseres Entdecken und Nutzen der persönlichkeitsrelevanten Hinweise zurückzuführen ist (Schritte 3 und 4 des RAM) und nicht darauf, dass sie durch ihr Verhalten mehr relevante Hinweise bei den Zielpersonen hervorrufen (Schritte 1 und 2 des RAM). Dies trifft laut Rogers und Biesanz zumindest auf kurze Interaktionen zu, wie sie in den vier Studien der Autoren verwendet wurden (ca. 3 Minuten).

2.3.4 Zusammenfassung und Fazit

Wie bereits zu Beginn angedeutet, existiert nicht der eine Ansatz zur Bestimmung der Akkuratheit von Persönlichkeitsbeurteilungen, sondern vielmehr eine gewisse Heterogenität in den Vorgehensweisen verschiedener Autor:innen. So besteht keine Einigkeit darüber, ob die Akkuratheit eher nach dem Ansatz der Profile Accuracy oder der Trait Accuracy bestimmt werden sollte (Back & Nestler, 2016). Die Entscheidung zwischen den Ansätzen scheint allerdings entscheidend zu sein, da die Gesamtwerte beider Ansätze keine große Übereinstimmung aufweisen (Hall et al., 2018) und möglicherweise verschiedene Fähigkeiten erfassen (Back & Nestler, 2016; Hall et al., 2018). Zudem muss bei Verwendung des Profile Accuracy-Ansatzes die dort enthaltene Stereotype Accuracy beziehungsweise Normative Accuracy-Komponente (Cronbach, 1955; Biesanz, 2010; Funder, 1999; Furr, 2008; Kenny & Winquist, 2001) bedacht und eventuell kontrolliert werden. Einerseits kann argumentiert werden, dass diese Komponente nicht mit einer Fähigkeit zu akkuraten individuellen Persönlichkeitsbeurteilungen in Verbindung gebracht werden kann, sondern eher allgemeines Wissen über Persönlichkeit widerspiegelt (Biesanz, 2010; Cronbach, 1955) – andererseits weisen Autor:innen darauf hin, dass bei der Kontrolle auch wahre Varianz entfernt wird, sofern das allgemeine Wissen auf die Zielperson zutrifft (Colvin & Bundick, 2001; Funder, 1999). Insgesamt erscheint von den vorgestellten Modellen und Methoden das SAM von Biesanz (2010) am umfassendsten, da es sowohl die Stereotype Accuracy-Komponente berücksichtigt als auch die Erkenntnis, dass es unterschiedlich gute Beurteiler:innen sowie unterschiedlich gute Zielpersonen gibt. Allerdings ist das SAM genau wie alle anderen Modelle und Methoden von einer sehr zentraleren Herausforderung betroffen: Die Bestimmung der wahren Persönlichkeitsausprägung der Zielperson, was nur über Kriterien erfolgen kann, die als Indikatoren der latenten Merkmalsausprägung dienen. Im Hinblick auf die Frage nach einer angemessenen Operationalisierung einer Fähigkeit zur akkuraten Beurteilung der Persönlichkeit anderer Personen kann daher der Nutzen aller vorgestellten Ansätze grundsätzlich in Frage gestellt werden. Dieser Aspekt wird in Kapitel 3 (Abschnitt 3.2) näher thematisiert.

3. Personality Understanding

Im Folgenden werden nacheinander zwei Fragen, die im Zusammenhang mit den zentralen Zielen der vorliegenden Arbeit stehen, thematisiert. Zum einen soll diskutiert werden, welche Fähigkeit akkuraten Beurteilungen der Persönlichkeit anderer Personen zu Grunde liegt und wie diese auf Basis der beschriebenen Theorien und Modellen konzeptualisiert und theoretisch eingeordnet werden kann. Im Anschluss daran wird die Frage nach einer angemessenen Operationalisierung dieser Fähigkeit adressiert. In diesem Zusammenhang soll auch diskutiert werden, inwieweit und ob die Bestimmung der Akkuratheit von Persönlichkeitsbeurteilungen hierfür in Frage kommt.

3.1 Konzeptualisierung

Einige der in Abschnitt 2.2 beschriebenen konzeptuellen Modelle und Theorien berücksichtigen eine Fähigkeit oder zumindest fähigkeitsrelevante Aspekte, die mit akkuraten Beurteilungen der Persönlichkeit anderer Personen in Verbindung gebracht werden können. Eine Übersicht über die Beschreibungen dieser Fähigkeiten oder Aspekte findet sich in Tabelle 3.1. Nicht berücksichtigt wurden in dieser Übersicht die ursprüngliche Version des Linsenmodells (Brunswik, 1956), das DLM (Hirschmüller et al., 2013), das STAM (Hall et al., 2017), das WAM (Kenny, 1991), das PERSON Modell (Kenny, 2004) sowie das TBM (West & Kenny, 2011), da hier keine Fähigkeiten beschrieben werden und auch keine anderen hierfür relevanten Aspekte, die nicht bereits durch andere Modelle abgedeckt werden. Die Interpersonale Intelligenz von Gardner (1991) sowie die PI von Mayer (2008) sind zwar in der Übersicht enthalten, müssen allerdings auf Grund der in den jeweiligen Abschnitten genannten Kritik mit Vorsicht betrachtet werden, sodass sie nicht in weitere Überlegungen miteinbezogen werden (vgl. auch Abschnitt 2.2.10).

Vergleich man die jeweiligen Fähigkeitsbeschreibungen, so fällt auf, dass diese zum Teil unterschiedliche kognitive Operationen beinhalten (z.B. Wissen vs. Verstehen). Eine vollständige Integration aller vorgestellten Modelle und Theorien ist somit nicht möglich und es muss diskutiert werden, welche der dort angenommenen Fähigkeiten für die akkurate Beurteilung individueller Personen am zentralsten ist, also den dort involvierten zentralen kognitiven Prozess am ehesten beschreiben könnte.

Tabelle 3.1

Übersicht über die Beschreibungen von Fähigkeiten und fähigkeitsrelevanten Aspekten mit Bezug zu akkuraten Beurteilungen der Persönlichkeit anderer Personen aus den in Abschnitt 2.2. vorgestellten Theorien und Modellen

Theorie/Modell	Name der Fähigkeit	Beschreibung der Fähigkeit
Realistic Accuracy Model (Funder, 1995)	-	Wissen über Persönlichkeit und wie sie sich im Verhalten manifestiert (Funder, 1995)
Dispositional Intelligence (Christiansen et al., 2005)	Dispositional Intelligence	„knowledge about personality and how it is related to behavior" (Christiansen et al., 2005, S. 124)
Attributionsmodelle (Gilbert et al., 1988; Trope, 1986)	-	Für akkurate Persönlichkeitsbeurteilungen müssen situative Einflüsse auf das Verhalten berücksichtigt werden.[b]
Realistic Accuracy Model (Funder, 1995)	Judgmental Ability	Fähigkeit zur valideren Nutzung entdeckter persönlichkeitsrelevanter Hinweise (Funder, 1995)
	[Teilfähigkeiten der Judgmental Ability][a]	Fähigkeit über die Relevanz der Hinweise zu entscheiden, die Hinweise angemessen zu gewichten, zu kombinieren sowie andere mögliche Ursachen für das Verhalten in Betracht zu ziehen (Letzring & Funder, 2021)
Structure-of-Intellect Model (Guilford, 1967)	Behavioral Cognition	„ability to understand the thoughts, feelings, and intentions of other people insofar as they are manifested in discernible behavior" (O'Sullivan & Guilford, 1975, S. 256)
Integratives Modell Sozialer Intelligenz (Weis & Süß, 2005)	Social Understanding	„Fähigkeit, gegebene soziale Informationen in der jeweiligen Situation zu verstehen und korrekt zu interpretieren" (Weis et al., 2006, S. 224)

Theorie/Modell	Name der Fähigkeit	Beschreibung der Fähigkeit
Kognitiv-Affektives Persönlichkeitssystem (Mischel & Shoda, 1995)	-	Für akkurate Persönlichkeitsbeurteilungen müssen die Wenn-Dann-Profile anderer Personen richtig konstruiert und angewendet werden (vgl. Kammrath et al., 2005; Mischel & Shoda, 1995).[b]
Theorie der Multiplen Intelligenzen (Gardner, 1991)	Interpersonal Intelligence	Fähigkeit zwischen den Stimmungen, Temperamenten, Motivationen und Bedürfnissen anderer Personen zu unterscheiden und angemessen auf diese zu reagieren (Gardner, 1991; Gardner & Hatch, 1989)
Personal Intelligence (Mayer, 2008)	Personal Intelligence	„capacity to reason about personality and to use personality and personal information to enhance one's thoughts, plans, and life experiences" (Mayer, 2008, S. 210)

Anmerkungen. [a] Zuordnung der genannten Teilfähigkeiten zu der Judgmental Ability geht nicht unmittelbar aus der Literatur hervor. [b] Das Modell schlägt keine Fähigkeit vor, thematisiert aber die hier zusammengefassten fähigkeitsrelevanten Aspekte.

Zwei der in Tabelle 3.1 berücksichtigten Theorien und Modelle nehmen eine Fähigkeit an, die als persönlichkeitsbezogenes Wissen beschrieben wurde (Funder, 1995; Christiansen et al., 2005), wobei die Konzeptualisierung von Christiansen et al. (2005) unter anderem auf der von Funder (1995) basiert. Bei der DI handelt es sich um ein eindeutig konzeptualisiertes sowie weitestgehend theoretisch fundiertes Konstrukt (vgl. Abschnitt 2.2.5). Wie zudem die Studien von Christiansen et al. (2005), De Kock et al. (2015) sowie Powell und Bourdage (2016) durch den Zusammenhang der DI mit akkuraten Persönlichkeitsbeurteilungen von Bekannten und Fremden aufzeigen konnten (Profile Accuracy:[2] r = .19 bis .42 für DI sowie r = .14 bis .33 für die drei Komponenten der DI), besitzt die DI eine gewisse Relevanz für die Akkuratheit von

[2] Zu beachten ist allerdings, dass die Zusammenhänge unter Verwendung des Trait Accuracy-Ansatzes (r = -.11 bis .08; Powell & Bourdage, 2016) sowie bei Bestimmung von Cronbachs (1955) Differential Accuracy (r = -.04 bis -.18, wobei das Vorzeichen auf Grund der Metrik der Differential Accuracy-Scores die intendierte Richtung aufweist; De Kock et al., 2015) geringer ausfielen.

Persönlichkeitsbeurteilungen. De Kock et al. (2020) identifizierten DI in einem systematischen Review über die Eigenschaften guter Beurteiler:innen im Personalmanagement, das auch eine metaanalytische Zusammenfassung der Ergebnisse beinhaltet, zudem als denjenigen Faktor, mit dem größten Zusammenhang zu akkuraten Persönlichkeitsbeurteilungen ($\bar{r}$ = .31). Hierbei ist anzumerken, dass in dem Review zwar prinzipiell nicht nur akkurate Beurteilungen der Persönlichkeit berücksichtigt wurden, in den Studien, die in den mittleren Zusammenhang zwischen DI und akkuraten Beurteilungen eingeflossen sind, allerdings ausschließlich die Persönlichkeit beurteilt wurde.

Trotz dieser Ergebnisse müssen auch die bereits in Abschnitt 2.2.5 angesprochenen, potentiell kritischen Aspekte bezüglich zwei der Komponenten der DI berücksichtigt werden: 1.) Trait Induction (vgl. Christiansen et al., 2005; De Kock et al., 2015): Inwieweit ist es für eine akkurate Beurteilung der Persönlichkeit einer individuellen Zielperson wichtig, Verhaltensweisen mit einer konkreten Persönlichkeitseigenschaft in Verbindung bringen zu können? 2.) Trait Extrapolation (vgl. Christiansen et al., 2005; De Kock et al., 2015): Wissen darüber, welche Persönlichkeitseigenschaften normalerweise gemeinsam bei einer Person auftreten, führt nur dann zu einer akkuraten Beurteilung, wenn die Zielperson diesem Profil auch entspricht. Nur in solchen Fällen kann allgemeines Wissen über Persönlichkeitseigenschaften fehlendes individuelles Wissen über die Zielperson ersetzen (De Kock et al., 2015; Funder, 1999; Paunonen & Hong, 2013; Vogt & Colvin, 2003). Es sei zudem angemerkt, dass die Trait Extrapolation konzeptuelle Überschneidungen mit der Stereotype Accuracy (Cronbach, 1955) beziehungsweise der Perceiver Normative Accuracy (Biesanz, 2010) aufweist. Diese Komponenten akkurater Urteile werden von den Autoren nicht mit einer Fähigkeit zu korrekten *individuellen* Persönlichkeitsbeurteilungen in Verbindung gebracht, sondern spiegeln eher allgemeines Wissen über Persönlichkeit wider (Biesanz, 2010; Cronbach, 1955) und werden zum Teil bei der Bestimmung der Akkuratheit auspartialisiert (Kenny & Winquist, 2001; Vogt & Colvin, 2003). Am wichtigsten erscheint die DI-Komponente Trait Contextualization. Diese bezieht sich auf die Verbindung zwischen Persönlichkeitseigenschaften und Situationen (Christiansen et al., 2005; De Kock et al., 2015), die wiederum auch in anderen Modellen thematisiert wird (Gilbert et al., 1988; Letzring & Funder, 2021; Trope, 1986). Insgesamt lässt sich sagen, dass die DI oder Komponenten der DI vermutlich dann eine wichtige Rolle bei der Beurteilung der Persönlichkeit anderer Personen spielen, wenn zu wenig relevante Information über die Zielperson vorliegt oder diese ein eher durchschnittliches Persönlichkeitsprofil aufweist (De Kock et al., 2015; Vogt & Colvin, 2003).

Auf Grund des Fokus auf normative Aspekte der Persönlichkeit lässt sich allerdings annehmen, dass DI nicht die zentrale Fähigkeit im Beurteilungsprozess darstellt.

Darüber hinaus ist festzuhalten, dass der in Tabelle 3.1 genannte fähigkeitsrelevante Aspekt der Attributionstheorien in den Ausführungen zum RAM von Letzring und Funder (2021) enthalten ist. Zudem stellt die Behaviorale Kognition aus dem SOI-Model (Guilford, 1967; O'Sullivan & Guilford, 1975) eine der Grundlagen für die Konzeptualisierung von SU im Integrativen Modell Sozialer Intelligenz dar (Weis & Süß, 2005) und wird damit durch SU abgedeckt. Übrig bleiben damit die Judgmental Ability aus dem RAM (Funder, 1995; Letzring & Funder, 2021), SU aus dem Integrativen Modell Sozialer Intelligenz (Weis & Süß, 2005; vgl. auch Weis, 2008) sowie die Wenn-Dann-Profile von Mischel und Shoda (1995; vgl. auch Kammrath et al., 2005). Diese weisen wiederum Gemeinsamkeiten auf, die eine Beschreibung des Prozesses der akkuraten Persönlichkeitsbeurteilung erlauben. So stimmen die drei Ansätze darin überein, dass für eine akkurate Beurteilung bereits vorhandene, oftmals behaviorale Hinweise auf die Persönlichkeitseigenschaft einer Zielperson korrekt interpretiert und verstanden werden müssen. Die Persönlichkeitseigenschaften äußern sich in der Regel in Form von typischen Verhaltensweisen der Zielperson, die zunächst als solche identifiziert werden müssen (Kammrath et al., 2005; Weis & Süß, 2005). Dies erfordert gegebenenfalls, dass einzelne Hinweise hinsichtlich ihrer Relevanz beurteilt, kombiniert, gewichtet und um situative Aspekte ergänzt werden (Letzring & Funder, 2021). Zudem müssen auf dessen Basis valide Schlussfolgerungen über die Persönlichkeitseigenschaften der Zielperson gezogen werden, beispielsweise in Form von Persönlichkeitsbeurteilungen oder Vorhersagen des typischen Verhaltens und Erlebens (Funder, 1995; Weis, 2008). Ausgehend von dieser Zusammenfassung und in Übereinstimmung mit der Annahme über die zentrale kognitive Anforderung bei SU (Weis, 2008), kann auch bei der Beurteilung der Persönlichkeit anderer Personen logisches Schlussfolgern als die zentrale kognitive Anforderung vermutet werden. Wie im Folgenden erläutert wird, liegen insbesondere Überschneidungen zum induktiven logischen Schlussfolgern vor.

Induktives logisches Schlussfolgern wurde von Schneider und McGrew (2018) definiert als „ability to observe a phenomenon and discover the underlying principles or rules that determine its behavior" (S. 93). Den Autoren zufolge geht es hierbei im Kern um das Erkennen und Verstehen von Mustern und Regelhaftigkeiten, was in Aufgaben zur Erfassung der Fähigkeit beispielsweise durch die Anwendung oder Angabe der Regelhaftigkeit gezeigt werden soll (Carroll, 1993; Schneider & McGrew, 2018). Bezogen auf den Bereich der Persönlichkeitsbeurteilungen können die Persönlichkeitseigenschaften in Form der typischen

Verhaltensweisen der Zielperson als die zu erkennenden Regelhaftigkeiten angesehen werden. Beispielsweise könnte aus der Beobachtung, dass eine Arbeitskollegin Unordnung auf ihrem Schreibtisch sowie in anderen Räumlichkeiten sofort aufräumt, geschlossen werden, dass die Kollegin typischerweise dazu neigt, Unordnung unmittelbar zu beseitigen. Hier wurde also eine Regelhaftigkeit im Verhalten der Kollegin auf Basis vorhandener Information erkannt, was nach der Definition von Schneider und McGrew (2018) induktivem logischen Schlussfolgern entspricht. Johnson-Laird (1994) definierte Induktion als „any process of thought yielding a conclusion that increases the semantic information in its initial observations or premise" (S. 11). Auch dieser Aspekt trifft auf den Prozess der Persönlichkeitsbeurteilung zu, da von spezifischer Information über Verhaltensweisen auf eine allgemeine Regelhaftigkeit geschlossen wird. Müssen die typischen Verhaltensweisen hingegen nicht erst identifiziert und verstanden werden, sondern sind diese bereits bekannt, ist eine Passung mit dem deduktiven logischen Schlussfolgern größer. Hierbei handelt es sich um „the ability to reason logically using known premises and principles" (Schneider & McGrew, 2018, S. 94) und es geht im Kern um die Anwendung der Regeln und das logische Ableiten korrekter Schlussfolgerungen (Carroll, 1993; Schneider & McGrew, 2018). Dies könnte zum Beispiel dann der Fall sein, wenn man Information über die typischen Verhaltensweisen einer Zielperson durch eine dritte Person, die die Zielperson sehr gut kennt, erhalten hat.

Es kann somit angenommen werden, dass logisches Schlussfolgern die zentrale kognitive Anforderung bei Beurteilungen der Persönlichkeit anderer Personen darstellt. Das vermutlich in erster Linie induktive logische Schlussfolgern findet hierbei in der Inhaltsdomäne Persönlichkeit statt, wobei der Fokus auf der Persönlichkeit anderer Personen liegt. Die so konzeptualisierte Fähigkeit wird in Analogie zu SU im Folgenden als *Personality Understanding* (PU) bezeichnet und soll in Anlehnung an Schneider und McGrew (2018) definiert werden als *die Fähigkeit, die Regelhaftigkeiten im Verhalten und Erleben anderer Personen, denen die Persönlichkeit der Personen zugrunde liegt, erkennen zu können*. Bei der Erfassung von PU kommt schließlich analog zum klassischen logischen Schlussfolgern die Anwendung oder Angabe der erkannten Regelhaftigkeiten, beispielsweise in Form von Persönlichkeitsbeurteilungen oder Verhaltensvorhersagen, hinzu.

Ausgehend von dieser Definition weist PU nicht nur hohe konzeptuelle Ähnlichkeit zu SU auf, sondern ebenfalls zu Emotional Understanding (EU). EU ist dem Four-Branch Model von Mayer und Salovey (1997; Mayer et al., 2016) zufolge eine Teilfähigkeit der Emotionalen Intelligenz (EI). In der aktuellsten Version des Modells wurde EU anhand verschiedener Bereiche des Schlussfolgerns beschrieben, die EU zugeordnet werden können (Mayer et al.,

2016). Hierzu gehören laut Mayer et al. (2016) unter anderem die zwei neu hinzugefügten Bereiche „understand how a person might feel in the future or under certain conditions" und „appraise the situations that are likely to elicit emotions" (S. 294). Nach Hellwig et al. (2020) handelt es sich bei den genannten Bereichen zwar um wichtige Beispiele für EU, den Autor:innen zufolge fehlt es der Beschreibung von EU im Four-Branch Model allerdings an einer Definition des den verschiedenen Bereichen zugrunde liegenden gemeinsamen kognitiven Prozesses. Eine vergleichbare Situation liegt bei den Theorien und Modellen aus dem Bereich akkurater Persönlichkeitsbeurteilungen vor. Hellwig et al. konzeptualisieren den Kern von EU in Form des zentralen kognitiven Prozesses als logisches Schlussfolgern im Inhaltsbereich Emotionen. Die Konzeptualisierung von PU wurde auf Grund der vorhandenen Parallelen daher auch in Anlehnung an diese Definition von EU vorgenommen. Darüber hinaus umfasst EU laut Hellwig et al. vermutlich auch eine zweite Komponente, bei der es sich um Wissen über typische emotionale Reaktionen handelt. Auch in diesem Punkt ist die Ähnlichkeit zum Bereich akkurater Persönlichkeitsbeurteilungen klar erkennbar. Hier existiert mit der DI ein Wissenskonstrukt (Christiansen et al., 2005; Funder, 1995), das die vorgeschlagene Konzeptualisierung von PU sinnvoll ergänzen könnte.

3.1.1 PU, SU und EU

PU weist somit konzeptuelle Überschneidungen sowohl mit SU als auch EU auf, sodass die Gemeinsamkeiten und Unterschiede der drei Konstrukte noch einmal näher betrachtet werden. Von besonderem Interesse ist hierbei die Frage, inwieweit PU und SU theoretisch voneinander abgrenzbar sind oder ob PU als Teilbereich von SU angesehen werden sollte.

3.1.1.1 Theoretische Überlegungen. Wie in Abschnitt 2.2.4 dargestellt, sind das Verstehen von Persönlichkeit und korrekte Persönlichkeitsbeurteilungen Bestandteil von Definitionen und Operationalisierungen der SI (O'Sullivan et al., 1965; Vernon, 1933; Wedeck, 1947). Auch im Integrativen Modell Sozialer Intelligenz werden Schlussfolgerungen über Persönlichkeitseigenschaften der Teilfähigkeit SU zugeordnet (Weis, 2008; Weis & Süß, 2005). Demzufolge könnte PU als Teilbereich von SU angesehen werden.

Dass Persönlichkeitseigenschaften bei der Konzeptualisierung von SU berücksichtigt wurden, ist leicht nachvollziehbar. Zum einen basiert das Integrative Modell unter anderem auf den zitierten Definitionen der SI, die Persönlichkeitsbeurteilungen als Inhalt dieser Fähigkeit annehmen (Weis, 2008; Weis & Süß, 2005). Zum anderen wird die Akkuratheit von Persönlichkeitsbeurteilungen traditionell zum Bereich Personenwahrnehmung oder Soziale

Kognition innerhalb der Sozialpsychologie gezählt und dort untersucht (Funder, 1995; Murphy, 2012). So wurde auch SU mit dem Konzept der Personenwahrnehmung aus der Sozialpsychologie verglichen und als äquivalent angesehen (Weis, 2008). Abgesehen davon, dass bei Schlussfolgerungen über die Persönlichkeit anderer Personen mindestens zwei Personen – Beurteiler:in und Zielperson – beteiligt sind, gibt es allerdings keinen Grund zur Annahme, dass es sich hierbei um eine Fähigkeit handelt, die vollständig als *soziale* Fähigkeit bezeichnet und somit vollständig SU zugeordnet werden kann. Dies ist vor allem dann der Fall, wenn man eine etwas engere Definition von SU zu Grunde legt. Laut Weis (2008) stellt die kognitive Anforderung von SU logisches Schlussfolgern dar, wobei soziale Information verarbeitet werden muss (Weis & Süß, 2005). Demzufolge kann SU auch als logisches Schlussfolgern in der sozialen Inhaltsdomäne verstanden werden (Schulze & Jobmann, 2016). Die entscheidende Frage ist hierbei allerdings, was unter sozialer Inhaltsdomäne verstanden werden kann und was hierzu gezählt werden sollte. Laut Weis (2008) umfasst diese Inhaltsdomäne soziale Stimuli wie Gesichtsausdrücke oder Interaktionen zwischen Personen, die wiederum Schlussfolgerungen über die Emotionen, Gedanken, Intentionen, Motivationen und Persönlichkeitseigenschaften einer Person erlauben sollen. Orientiert man sich wiederum an der klassischen Definition von Sozialpsychologie als „an attempt to understand and explain how the thought, feeling, and behavior of individuals are influenced by the actual, imagined, or implied presence of others" (Allport, 1985, S. 3), so liegt ein Fokus der sozialen Inhaltsdomäne auf Interaktionen zwischen Personen und Schlussfolgerungen über beispielsweise Beziehungen zwischen Personen oder Gruppen sowie Verhalten auf Gruppenebene näher. Dies hätte eine engere Definition von SU zur Folge, da beispielsweise nicht alle Emotionen (z.B. Angst auf Grund situativer Aspekte) oder Persönlichkeitseigenschaften (Neurotizismus, Offenheit für Erfahrungen, Gewissenhaftigkeit; McCrae & Costa, 1989) zwingend die (vorgestellte oder implizierte) Anwesenheit anderer Personen, das heißt eine soziale Interaktion, erfordern.

Legt man diese engere Definition von SU zu Grunde, so stellt PU zwar keinen Teilbereich von SU dar, es kann aber angenommen werden, dass sich beide Fähigkeiten substantiell überschneiden. Dieser Überschneidungsbereich könnte beschrieben werden als logisches Schlussfolgern über Persönlichkeitseigenschaften, die sich in sozialen Interaktionen zeigen und hier eine wichtige Rolle spielen. Dies trifft auf die Big Five-Faktoren Extraversion und Verträglichkeit zu, von denen angenommen wird, dass sie für soziale Interaktionen und interpersonales Verhalten eine wichtige Rolle spielen (McCrae & Costa, 1989). John and Srivastava (1999) schlugen kurze Definitionen der Big Five vor und betonten hierdurch bei der

Extraversion den „energetic approach to the social [...] world“ (S. 121) und bei der Verträglichkeit die „prosocial and communal orientation toward others“ (S. 121), was die soziale Orientierung der beiden Faktoren verdeutlicht. Spezifischere Beispiele finden sich im FFM: Die Extraversions-Facette Herzlichkeit beschreibt unter anderem die Tendenz, „leicht enge Bindungen zu anderen“ einzugehen (S. 41) und weist eine gewisse Wichtigkeit für Beziehungen zwischen Personen auf (Ostendorf & Angleitner, 2004). Ein anderes Beispiel für die Extraversion ist die Facette Geselligkeit, die nach Ostendorf und Angleitner (2004) die Tendenz beschreibt, gerne mit vielen Personen Zeit zu verbringen. Watson und Clark (1997) schlugen ein integratives Modell der Extraversion vor, in dem die beiden zuvor genannten Facetten diejenige Komponente der Extraversion repräsentieren, bei der interpersonale Beziehungen und Interaktionen im Fokus stehen. Auch für den Faktor Verträglichkeit finden sich im FFM spezifischere Beispiele. So wird die Facette Altruismus als „aktive Besorgnis um das Wohlergehen anderer“ (S. 45) beschrieben und die Facette Entgegenkommen steht für interindividuelle Unterschiede im Verhalten bei Konflikten mit anderen Personen (Ostendorf & Angleitner, 2004). Besonders deutlich wird der soziale Charakter bei Graziano und Tobin (2018), denen zufolge Verträglichkeit als Motivation, positive soziale Beziehungen herzustellen und aufrechtzuerhalten, beschrieben werden kann.

In gleicher Weise lässt sich ein Überschneidungsbereich zwischen PU und EU beschreiben, der das logische Schlussfolgern über Persönlichkeitseigenschaften beinhaltet, bei denen Emotionen einen wichtigen Bestandteil darstellen. Hier stehen somit Emotionen im Fokus, die eine typische Reaktion der Zielperson darstellen. Dies ist vor allem bei dem Big Five-Faktor Neurotizismus der Fall (McCrae & John, 1992), der durch negative Emotionalität an dem einen und emotionale Stabilität sowie Gelassenheit am anderen Ende charakterisiert werden kann (John & Srivastava, 1999). Als spezifischere Beispiele können die Facetten Ängstlichkeit und Depression aus dem FFM herangezogen werden, die die Tendenz zum Erleben von Besorgnis, Furcht und Nervosität beziehungsweise Schuld, Traurigkeit und Hoffnungslosigkeit darstellen (Ostendorf & Angleitner, 2004). Aber auch Teile der Extraversion können in den Überschneidungsbereich zwischen PU und EU eingeordnet werden. So stellen Watson und Clark (1997) zufolge interindividuelle Unterschiede in positiver Emotionalität den Kern von Extraversion dar (vgl. auch Facette Frohsinn aus dem FFM; Ostendorf & Angleitner, 2004). Die Annahme, dass PU und EU konzeptuell zusammenhängen, passt darüber hinaus auch zu den Annahmen des STAM von Hall et al. (2017; vgl. Abschnitt 2.2.3), in dem eine kausale Beziehung zwischen akkuraten Urteilen über affektive Zustände und akkuraten Urteilen über Persönlichkeitseigenschaften beschrieben wird.

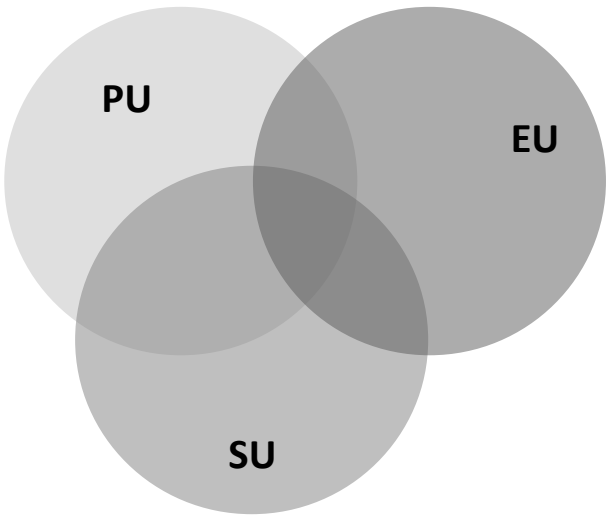
Auch der Überschneidungsbereich zwischen SU und EU ließe sich auf ähnliche Weise als logisches Schlussfolgern über Emotionen, die sich in sozialen Interaktionen zeigen und hier eine wichtige Rolle spielen, beschreiben. Beispiele wären die Emotionen Liebe und Abneigung, die durch andere Personen ausgelöst werden (Roseman, 2001). Der Überschneidungsbereich steht allerdings nicht im Fokus der Arbeit und wird daher nicht näher betrachtet. Eine Diskussion der Gemeinsamkeiten und Unterschiede zwischen den breiteren Fähigkeiten EI und SI sowie diesbezügliche offene Fragen finden sich beispielsweise bei Mayer et al. (2016), Kang et al. (2005) sowie Weis und Süß (2005).

Letztendlich lässt sich auch ein Überschneidungsbereich aller drei Konstrukte annehmen, in dem Persönlichkeitseigenschaften, Emotionen und soziale Interaktionen gemeinsam Bestandteil des logischen Schlussfolgerns sind. Dies könnte dann der Fall sein, wenn es um Persönlichkeitseigenschaften geht, die sich in sozialen Interaktionen zeigen und bei denen Emotionen einen wichtigen Bestandteil darstellen. Ein mögliches Beispiel ist die Facette Soziale Befangenheit, die im FFM dem Neurotizismus zugeordnet wird (Ostendorf & Angleitner, 2004). Nach Ostendorf und Angleitner (2004) umfasst diese Facette unter anderem die Tendenz, sich auf Grund der Anwesenheit anderer Personen nicht wohl zu fühlen sowie das Erleben der Emotionen Scham und Verlegenheit.

Eine grafische Veranschaulichung der angenommenen Überschneidungsbereiche der drei Fähigkeiten findet sich in Abbildung 3.1.

Abbildung 3.1

Grafische Veranschaulichung der angenommenen Überschneidungsbereiche von Personality Understanding (PU), Social Understanding (SU) und Emotional Understanding (EU)



Zusammenfassend lässt sich sagen, dass die Gemeinsamkeit der drei Konstrukte PU, SU und EU in der zentralen kognitiven Operation (logisches Schlussfolgern) und die Unterschiede in der inhaltlichen Verankerung (Persönlichkeit vs. interpersonale Interaktionen vs. Emotionen) liegen, wobei bei der inhaltlichen Verankerung ebenfalls Überschneidungen angenommen werden können. Aus theoretischer Sicht lassen sich daher positive Beziehungen der drei Konstrukte vermuten.

Die angenommenen Gemeinsamkeiten und Unterschiede zwischen PU, SU und EU stehen im Einklang mit den Ansichten von Mayer et al. (2016), die PI, SI und EI auf eine ähnliche Weise verglichen haben (vgl. auch Mayer, 2009). Laut Mayer et al. (2016) können die drei Konstrukte als Schlussfolgern in unterschiedlich komplexen Bereichen (Emotionen, Persönlichkeit, soziale Prozesse) charakterisiert werden. Auch wenn sich die Autor:innen auf die PI, die bei der Ableitung von PU nicht berücksichtigt wurde, sowie die EI des Four-Branch Models beziehen, passt dies zum vorgenommenen Vergleich zwischen PU, SU und EU. Das liegt unter anderem daran, dass PU – obwohl nicht intendiert – konzeptuelle Gemeinsamkeiten mit dem zweiten Fähigkeitsbereich der PI aufweist, der unter anderem das Bilden von Modellen über die Persönlichkeit anderer Personen umfasst (Mayer, 2008, 2009). Zudem legten auch Mayer et al. (2016) dem Vergleich eine Definition von SI als Schlussfolgern über Gruppen und Beziehungen zwischen Personen und Gruppen zu Grunde, die somit weitestgehend der oben beschriebenen engeren Beschreibung von SU entspricht.

3.1.1.2 Empirische Evidenz. Empirische Evidenz für die theoretisch angenommenen Zusammenhänge zwischen PU und SU sowie EU ist auf Grund der neuen Konzeptualisierung der Fähigkeit PU und des Fokus bei allen drei Konstrukten auf das logische Schlussfolgern nicht vorhanden. Es werden daher alternativ Studien herangezogen, die Zusammenhänge zwischen konzeptuell ähnlichen Konstrukten (z.B. EI und PI) und Methoden (z.B. die Bestimmung der Akkuratheit von Persönlichkeitsbeurteilungen) betrachtet haben.

In einer Studie von Mayer et al. (2012) ergab sich für den Zusammenhang zwischen PI (gemessen mit dem TOPI, Version 1.2) und EU (gemessen mit dem Mayer-Salovey-Caruso Emotional Intelligence Test [MSCEIT], einem Leistungstest zur Erfassung der vier EI-Teilbereiche des Four-Branch Models; Mayer et al., 2003) ein großer Effekt von $r = .68$, $p < .01$. Die Studien von Schlegel et al. (2017) sowie Jaksic und Schlegel (2020) lieferten hingegen andere Ergebnisse. Schlegel et al. (2017) untersuchten in einer Meta-Analyse, wie hoch Tests zur Erfassung akkurater Beurteilungen in verschiedenen Inhaltsbereichen zusammenhängen. Für die beiden Inhaltsbereiche Persönlichkeit und Emotionen ergab sich eine durchschnittliche

Korrelation von $\bar{r}$ = .09, p < .01 (Schlegel et al., 2017). Zusätzlich zu den von Schlegel et al. (2017) diskutierten möglichen methodischen Gründen für diese sehr geringe durchschnittliche Korrelation (u.a. geringe Reliabilität, unterschiedliche Antwortformate) muss hierbei beachtet werden, dass weder berücksichtigt noch angegeben wurde, wie die akkuraten Persönlichkeitsbeurteilungen in den verwendeten Studien bestimmt wurden, sowie dass die akkuraten Persönlichkeitsbeurteilungen untereinander eine ähnlich geringe Korrelation aufwiesen ($\bar{r}$ = .08, p < .05, vermutlich Trait Accuracy; vgl. Hall et al., 2018). Unterschiedliche Methoden zur Bestimmung der Akkuratheit korrelieren den Ergebnissen von Hall et al. (2018) zufolge maximal klein bis mittel miteinander, was unterstreicht, dass die Ergebnisse von Schlegel et al. (2017) mit Vorsicht interpretiert werden müssen. Hinzu kommt, dass die in den Analysen für den Inhaltsbereich Emotionen berücksichtigten Studien hauptsächlich Testverfahren zur Erfassung der Emotionserkennungsfähigkeit verwendet haben (vgl. Jaksic & Schlegel, 2020; Schlegel et al., 2017), was nach dem Four-Branch Model (Mayer & Salovey, 1997; Mayer et al., 2016) einen anderen Teilbereich der EI darstellt und nicht mit EU gleichgesetzt werden kann. Aus diesen Gründen erscheint eine Übertragung auf den Zusammenhang zwischen PU und EU weniger angemessen. Jaksic und Schlegel (2020) untersuchten nach eigenen Angaben erstmals den Zusammenhang zwischen akkuraten Persönlichkeitsbeurteilungen und EU. Für die Bestimmung der Akkuratheit verwendeten die Autorinnen unterschiedliche Methoden (Trait Accuracy, Profile Accuracy, Distinctive Accuracy) und EU wurde mit Hilfe des Situational Test of Emotion Understanding (STEU) von MacCann und Roberts (2008) erfasst, einem Leistungstest zur Erfassung dieser einen Teilfähigkeit der EI nach dem Four-Branch Model. Es ergaben sich Korrelationen zwischen r_s = .03 (n.s.; Distinctive Accuracy) und .08 (n.s.; Trait Accuracy), wobei beachtet werden muss, dass die Akkuratheitswerte sehr geringe Reliabilitäten von maximal .36 aufwiesen (Jaksic & Schlegel, 2020). Jaksic und Schlegel (2020) untersuchten allerdings ebenfalls den Zusammenhang zwischen akkuraten Persönlichkeitsbeurteilungen und der Emotionserkennungsfähigkeit. Hier ergaben sich mit r_s = .11 (n.s.; Profile Accuracy), .21 (p < .05; Distinctive Accuracy) und .33 (p < .001; Trait Accuracy) zum Teil deutlich größere Zusammenhänge.

Aus den Ergebnissen der drei zitierten Studien lässt sich zusammenfassend schließen, dass es offensichtlich von der konkreten Konzeptualisierung und Operationalisierung abhängt, ob Zusammenhänge zwischen akkuraten Persönlichkeitsbeurteilungen, beziehungsweise einer angenommenen zugrunde liegenden Fähigkeit, und Teilbereichen der EI erwartet und gefunden

werden können. Eine Übertragung auf den angenommenen Zusammenhang zwischen PU und EU ist daher maximal mit Einschränkungen möglich.

Eine für den angenommenen Zusammenhang zwischen PU und SU interessante Studie führten Speer et al. (2019) durch. Die Autor:innen entwickelten einen Situational Judgment Test zur Erfassung der SI auf Basis einer eigenen Definition des Konstrukts, die sie durch Kombination verschiedener Elemente vorhandener Definitionen und Konzeptualisierungen (u.a. Weis & Süß, 2005) ableiteten. Im Rahmen von drei Studien untersuchten Speer et al. (2019) den Zusammenhang von SI zu EI (r = .28, p < .01; erfasst mit dem MSCEIT), DI (r = .51, p < .01; erfasst mit dem Test von Christiansen et al., 2005) sowie akkuraten Persönlichkeitsbeurteilungen (Differential Accuracy nach Cronbach, 1955: r = .29, p < .01, wobei die Distanzwerte invertiert wurden; Profile Accuracy: r = .24, p < .01). Die Übertragbarkeit auf den Zusammenhang zwischen PU und SU ist allerdings auch hier nur eingeschränkt möglich, insbesondere auf Grund der gewählten Operationalisierung von SI. In dem neu konstruierten Test wurden den Testpersonen arbeitsbezogene Situationen präsentiert, die einen interpersonalen Konflikt oder eine interpersonale Anforderung beinhalteten (Speer et al., 2019). Im Anschluss wurden vier mögliche Reaktionen vorgeschlagen, von denen die am meisten und die am wenigstens effektive ausgewählt werden mussten. Während die SI-Definition der Autor:innen noch Überschneidungen mit SU aufweist, ist dies in der Operationalisierung nicht mehr der Fall, da der Fokus – so wie er aus den bei Speer et al. präsentierten Beispielaufgaben hervorgeht – eher auf der behavioralen Komponente der SI liegt anstatt auf der kognitiven, wie es bei SU der Fall ist (Weis & Süß, 2005). Auffällig ist zudem, dass sich in der Studie von Speer et al. ein größerer Zusammenhang zwischen SI und DI ergab als zwischen SI und akkuraten Persönlichkeitsbeurteilungen, was darauf hindeutet, dass der SI-Test mehr mit interpersonalem Wissen zu tun haben könnte.

Letztlich soll noch einmal daran erinnert werden, dass Persönlichkeitsbeurteilungen auch Bestandteil der Operationalisierung von SU im MTSI sind (Süß et al., 2009; Weis, 2008; vgl. Abschnitt 2.2.4.2). Diese werden bei der Auswertung allerdings nicht berücksichtigt (Süß et al., 2009) und es zeigte sich in konfirmatorischen Faktorenanalysen, dass die Einbeziehung der Persönlichkeitsbeurteilungen einen schlechteren Modellfit der Aufgaben zur Erfassung von SU sowie Sozialem Gedächtnis zur Folge hatte (Seidel, 2007). Dies ist ein erster, wenn auch sehr vager Hinweis darauf, dass akkurate Persönlichkeitsbeurteilungen nicht vollständig der SU zugeordnet werden können.

Alles in allem lässt sich sagen, dass die empirische Evidenz in Abhängigkeit der verwendeten Instrumente variabel ausfällt sowie die Übertragbarkeit auf die Zusammenhänge

zwischen PU und SU sowie PU und EU nur äußerst eingeschränkt möglich ist. Die zitierten Studien können im Hinblick auf die theoretisch angenommenen Zusammenhänge der drei Konstrukte somit maximal erste Anhaltspunkte liefern.

3.1.2 PU und Intelligenz

3.1.2.1 Theoretische Überlegungen. PU wird in der vorliegenden Arbeit als Fähigkeit zum logischem Schlussfolgern im Inhaltsbereich Persönlichkeit konzeptualisiert. Als kognitive Operation wird somit eine klassische und zentrale Fähigkeit aus dem Intelligenzbereich angenommen. Aus diesem Grund soll PU als nächstes im Kontext bekannter Intelligenzmodelle betrachtet werden. Hierfür werden das Berliner Intelligenzstrukturmodell (BIS-Modell; Jäger, 1982, 1984) sowie die Cattell-Horn-Carroll Theorie kognitiver Fähigkeiten (CHC-Theorie; McGrew, 2009) herangezogen, da beide Ansätze offen für die Integration weiterer Forschung sind (Jäger, 1984; McGrew, 2009). Zudem wurden hier, wie im Folgenden näher beschrieben wird, bereits die mit PU konzeptuell verwandten Konstrukte SI und EI eingeordnet.

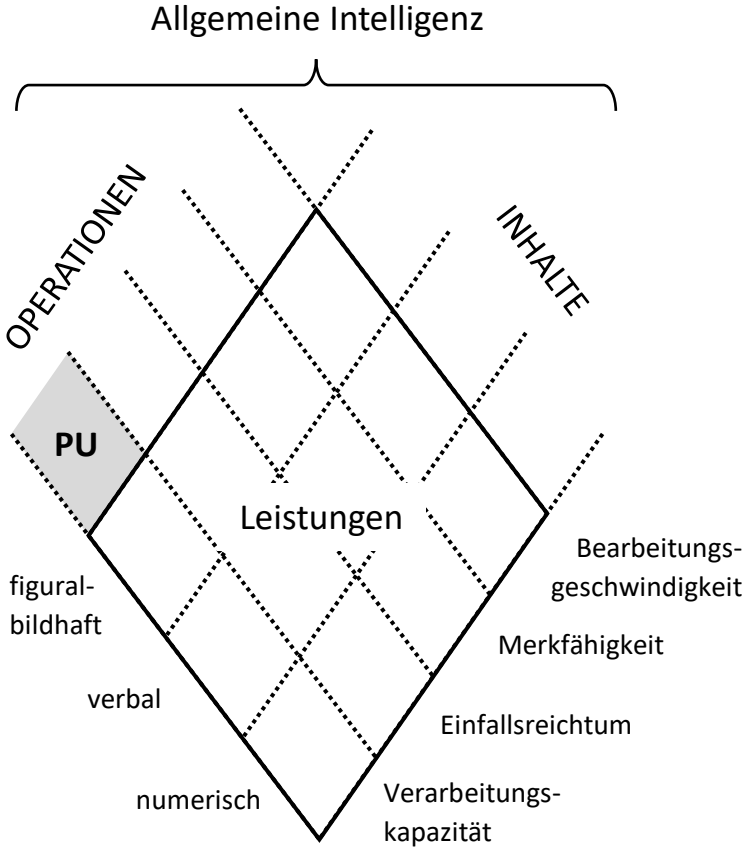
Beim BIS-Modell von Jäger (1982, 1984; vgl. auch Jäger et al., 1997) handelt es sich um ein integratives und hierarchisches Intelligenzstrukturmodell, in dem angenommen wird, dass Intelligenzleistungen anhand der zwei Facetten *Operationen* und *Inhalte* gruppiert werden können. Diese beiden Facetten spiegeln die zur Aufgabenlösung notwendige kognitive Operation sowie den Inhalt des zu verarbeitenden Aufgabenmaterials wider. Auf Seite der Facette Operationen unterscheidet das Modell in der aktuellsten Version (Jäger et al., 1997) die vier Fähigkeiten Verarbeitungskapazität, Einfallsreichtum, Merkfähigkeit und Bearbeitungsgeschwindigkeit und auf Seite der Inhalte die drei Fähigkeiten Sprachgebundenes Denken (verbal), Zahlengebundenes Denken (numerisch) und Anschauungsgebundenes Denken (figural-bildhaft). Eine vollständige Kreuzung der vier Operationen mit den drei Inhalten resultiert in zwölf Intelligenzleistungen, die die einzelnen Zellen des Modells darstellen (vgl. Abbildung 3.2) und die jeweils im Wesentlichen durch eine operative und eine inhaltsgebundene Fähigkeit bedingt sind (Jäger, 1982, Jäger et al., 1997). Auf hierarchisch höherer Ebene findet sich im BIS-Modell zudem die *Allgemeine Intelligenz*, die als „Integral aller Fähigkeiten“ (S. 4) verstanden wird (Jäger et al., 1997; vgl. auch Jäger, 1982, 1984).

Jäger betonte im Zusammenhang mit seinem Modell mehrfach (z.B. Jäger, 1982, 1984; Jäger et al., 1997), dass es offen für Erweiterungen ist und bei theoretischer und empirischer

Fundierung weitere Operationen und Inhalte ergänzt werden können. Er selbst schlug unter anderem die Ergänzung eines sozialen Inhaltes vor (Jäger, 1982).

Abbildung 3.2

Berliner Intelligenzstrukturmodell (BIS-Modell) von Jäger (1982, 1984) und Einordnung des Konstruktes Personality Understanding (PU)



Anmerkung. Abbildung nach Jäger et al. (1997).

Das Integrative Modell Sozialer Intelligenz unterscheidet in Anlehnung an das BIS-Modell ebenfalls eine operative Facette und eine Inhaltsfacette (Weis et al., 2006). Eine explizite Einordnung von Teilfähigkeiten der SI in das BIS-Modell nahm zudem Weis (2008) vor. So kann der Autorin zufolge SU durch Kreuzung der Operation Verarbeitungskapazität mit einem neuen sozialen Inhalt als neue Zelle in das BIS-Modell integriert werden. Eine analoge Einordung von EU, konzeptualisiert als logisches Schlussfolgern in der Inhaltsdomäne Emotionen (Hellwig, et al., 2020), durch die Kreuzung der Operation Verarbeitungskapazität mit einem neuen emotionalen Inhalt wurde ebenfalls bereits vorgenommen (Schulze & Jobmann, 2016). Verarbeitungskapazität wird im BIS-Modell definiert als „Verarbeitung

komplexer Informationen bei Aufgaben, die nicht auf Anhieb zu lösen sind, sondern Heranziehen, vielfaltiges Beziehungsstiften, formallogisch exaktes Denken und sachgerechtes Beurteilen von Informationen erfordern“ (Jäger et al., 1997, S. 6) und entspricht im Wesentlichen der Fähigkeit zum logischen Schlussfolgern (Beauducel & Kersting, 2002; Süß et al., 2002). PU kann somit analog zu SU und EU durch Kreuzung der Operation Verarbeitungskapazität mit dem neuen Inhaltsbereich Persönlichkeit in das BIS-Modell eingeordnet werden (vgl. Abbildung 3.2). Es sei an dieser Stelle allerdings darauf hingewiesen, dass die in den Zellen des BIS-Modells darstellten Intelligenzleistungen bisher nicht als Fähigkeiten konzipiert wurden (Jäger et al., 1997). Da laut Jäger et al. (1997) das Modell aber dazu genutzt werden kann, um Aufgaben und Tests zu organisieren, erscheint eine Einordnung von PU sinnvoll, um einen Orientierungsrahmen für spätere Validierungsstudien von Aufgaben zur Erfassung von PU abzuleiten (insbesondere bezüglich der konvergenten und diskriminanten Validität). Zudem gab es mittlerweile erste Vorschläge, die einzelnen Zellen ebenfalls als Fähigkeiten zu interpretieren (Jäger et al., 2006).

Die CHC-Theorie stellt dem Namen entsprechend eine Vereinigung der Gf-Gc Theorie von Cattell und Horn (Horn & Cattell, 1966; Horn & Noll, 1997) sowie der Three-Stratum Theorie von Carroll (1993) dar (McGrew, 2009). Sie soll eine umfassende Taxonomie kognitiver Fähigkeiten darstellen und die Einordnung von Studienergebnissen erleichtern (Schneider & McGrew, 2018). Die CHC-Theorie weist eine hierarchische Organisation von Fähigkeitsdimensionen unterschiedlicher Breite auf (Schneider & McGrew, 2018). Unterschieden werden insbesondere *Narrow Abilities*, die auch als Stratum I-Fähigkeiten bezeichnet werden und Interkorrelation mehrerer spezifischer kognitiver Leistungen erklären sollen, und die hierarchisch höheren *Broad Abilities* (Stratum II-Fähigkeiten), die wiederum mehrere interkorrelierte Stratum I-Fähigkeiten umfassen (McGrew, 2009; Schneider & McGrew, 2018). Beispielsweise werden fluides logisches Schlussfolgern (fluide Intelligenz; Gf), Wissen (kristalline Intelligenz; Gc) und Verarbeitungsgeschwindigkeit (Gs) in die Ebene der Stratum II-Fähigkeiten eingeordnet (McGrew, 2009). Unterhalb von Gf befinden sich unter anderem die Stratum I-Fähigkeiten induktives logisches Schlussfolgern, deduktives logisches Schlussfolgern und quantitatives logisches Schlussfolgern (McGrew, 2009).

Auch die CHC-Theorie ist offen für Erweiterungen und Überarbeitungen, die bereits vorgenommen wurden und auch weiterhin werden (McGrew, 2009; Schneider & McGrew, 2018). Im Rahmen der neusten Aktualisierung sprechen sich Schneider und McGrew (2018) für die Aufnahme von EI in die CHC-Theorie als vorläufige Stratum II-Fähigkeit aus. Eine hierfür ausschlaggebende Studie, in der die faktorielle Struktur mehrerer Testverfahren zur

Erfassung von fünf Stratum II-Fähigkeiten sowie zur Erfassung von drei Teilfähigkeiten der EI untersucht wurde, stammt von MacCann et al. (2014; vgl. Schneider & McGrew, 2018). Die Ergebnisse sprechen den Autor:innen zufolge dafür, dass EI als zusätzliche Stratum II-Fähigkeit in die CHC-Theorie aufgenommen werden kann (MacCann et al., 2014). Die in der Studie erfassten Teilfähigkeiten der EI, wozu auch EU gehörte, schlugen MacCann et al. (2014) als Stratum I-Fähigkeiten vor. Evans et al. (2020) konnten diese Ergebnisse unter Verwendung anderer Operationalisierungen weitestgehend replizieren und somit die Schlussfolgerung von MacCann et al. (2014), dass EI eine Stratum II-Fähigkeit mit drei untergeordneten Stratum I-Fähigkeiten (u.a. EU) darstellt, stützen. PU ist als spezifische Form des logischen Schlussfolgerns ebenfalls auf Ebene der Stratum I-Fähigkeiten einzuordnen. Unklar ist allerdings, unter welcher breiteren Stratum II-Fähigkeit PU eingeordnet werden kann. Hier sind insbesondere zwei Einordnungen denkbar: 1.) Unterhalb von Gf ist mit dem quantitativen logischen Schlussfolgern bereits eine spezifischere Form des logischen Schlussfolgerns eingeordnet (McGrew, 2009). Hierbei handelt es sich um die Fähigkeit, mit Mengen, mathematischen Beziehungen und Operatoren zu schlussfolgern (Schneider & McGrew, 2018, S. 94), also um das logische Schlussfolgern im mathematischen Inhaltsbereich. PU könnte somit analog hierzu unterhalb von Gf eingeordnet werden. 2.) Alternativ wäre vorstellbar, dass PU, ähnlich wie EU, einer bisher nicht identifizierten breiteren Stratum II-Fähigkeit untergeordnet werden kann, die sämtliche kognitive Fähigkeiten im Inhaltsbereich Persönlichkeit zusammenfasst (u.a. PU und DI).

3.1.2.2 Empirische Evidenz. Der Zusammenhang zwischen akkuraten Persönlichkeitsbeurteilungen und Intelligenz – allerdings nicht logischem Schlussfolgern im Speziellen – wurde bereits sehr früh in der Forschung untersucht. Im Folgenden werden zunächst die Ergebnisse aus vier Reviews und Meta-Analysen berichtet, die sich mit dem Zusammenhang zwischen akkurater Personenwahrnehmung und Intelligenz beschäftigt haben. Diese Arbeiten haben sich allerdings nicht auf Studien zu akkuraten Beurteilungen der Persönlichkeit anderer Personen beschränkt, sondern beispielsweise auch solche zu akkuraten Beurteilungen von Emotionen berücksichtigt. Dies muss bei der Interpretation der Studienergebnisse berücksichtigt werden.

In seinem vielfach zitierten Review kommt Taft (1955) zu dem Schluss, dass zwischen der Fähigkeit zum akkuraten Beurteilen anderer Personen und Intelligenz ein positiver Zusammenhang besteht. Er berücksichtigte neben akkuraten Beurteilungen der Persönlichkeit auch solche von beispielsweise Fähigkeiten oder Emotionen. Unter anderem bezog sich Taft

(1955) auf die Studie von Vernon (1933), der zwischen akkuraten Beurteilungen der Persönlichkeit von Fremden und Intelligenz einen Zusammenhang von $r = .31$ fand.

Davis und Kraus (1997) ermittelten in ihrer Meta-Analyse zwischen interpersonaler Akkuratheit und intellektuellen Fähigkeiten (Intelligenz und andere Leistungsmaße) einen mittleren Zusammenhang von $\bar{r} = .23$, $p < .001$, was einen der größten Effekte ihrer Analyse darstellte. Unter interpersonaler Akkuratheit fassten der Autor und die Autorin neben akkuraten Persönlichkeitsbeurteilungen unter anderem auch die akkurate Identifikation von Emotionen und anderen nicht-affektiven Zuständen oder auch die akkurate Einschätzung von Identität und Status der Zielpersonen zusammen.

Murphy und Hall (2011) kritisierten die Meta-Analyse von Davis und Kraus (1997) dahingehend, dass diese unvollständig sei, da nicht alle zu dieser Zeit verfügbaren Studien berücksichtigt wurden. Sie führten darauf aufbauend eine eigene Meta-Analyse durch und ermittelten für den Zusammenhang zwischen interpersonaler Sensitivität und Intelligenz einen mittleren Effekt von $\bar{r} = .19$, $p < .001$ (Murphy & Hall, 2011). Bei der Interpretation dieses Ergebnisses ist zu berücksichtigen, dass die Autorinnen nach eigener Angabe Intelligenz eher breit operationalisierten und neben standardisierten Intelligenztests auch andere leistungsbezogene Maße wie Schulleistungstests und Schulnoten berücksichtigten. Zudem stellten Studien zu akkuraten Persönlichkeitsbeurteilungen in dieser Meta-Analyse eher die Ausnahme dar. Unter interpersonaler Sensitivität verstanden Murphy und Hall (2011) die Fähigkeit „to accurately detect the state(s) or trait(s) of unacquainted others“ (S. 56) und betonten dabei, dass es hier vor allem um die akkurate Wahrnehmung interpersonaler Information geht und nicht um das Verständnis oder die Interpretation, was somit nicht zu dem Fokus von PU passt. Eine von den Autorinnen erstellte Übersichtstabelle (Murphy & Hall, 2011, Tabelle 1) ordnet zudem nur eine Studie ausschließlich dem Bereich akkurater Persönlichkeitsbeurteilungen zu (Lippa & Dietz, 2000), obwohl zumindest eine weitere Studie in der Tabelle enthalten ist, die in diesen Bereich passt (Christiansen et al., 2005).

Darüber hinaus soll noch einmal auf das bereits zitierte systematische Review über die Eigenschaften guter Beurteiler:innen im Personalmanagement von De Kock et al. (2020) eingegangen werden. Hier zeigte sich für allgemeine Intelligenz mit $\bar{r} = .18$ der im Mittel zweitgrößte Zusammenhang mit akkuraten Personenbeurteilungen. In diesen Zusammenhang flossen erneut nicht ausschließlich Studien ein, die sich mit akkuraten Persönlichkeitsbeurteilungen beschäftigt haben, sondern auch solche, bei denen beispielsweise die Beurteilung der (Arbeits-)Leistung untersucht wurde. Wie von De Kock et al. (2020)

angemerkt wurde, zeigen die im Review berücksichtigten Studien große Unterschiede hinsichtlich der Höhe des Zusammenhangs. Dies ist auch dann noch der Fall, wenn man nur diejenigen Studien betrachtet, bei denen akkurate Persönlichkeitsbeurteilungen untersucht wurden. So ergab sich in der Studie von Letzring (2008) mit r = -.01, p = .89 kein Zusammenhang zwischen allgemeiner Intelligenz und akkuraten Persönlichkeitsbeurteilungen (Profile Accuracy). Bei De Kock et al. (2015) zeigten sich Zusammenhänge von r = - .12, n.s. (Differential Accuracy nach Cronbach, 1955; negatives Vorzeichen ist in intendierter Richtung) beziehungsweise r = .20, p < .05 (Profile Accuracy) und bei Christiansen et al. (2005) Zusammenhänge von r = .13, n.s. (Multiple-Choice Test zur Erfassung akkurater Persönlichkeitsbeurteilungen auf Basis standardisierter Job-Interviews) beziehungswiese r = .24, p < .01 (Profile Accuracy). In der Studie von Lippa und Dietz (2000) resultierte mit r = .36, p < .01 (Trait Accuracy) der deskriptiv größte Zusammenhang. Interessant ist im Zusammenhang mit den Ergebnissen dieser Studien zudem, dass in allen Studien der gleiche 12-minütige Intelligenztest, allerdings in unterschiedlichen Versionen, verwendet wurde.

Ergänzt werden können die bisher präsentierten Ergebnisse zum Zusammenhang zwischen allgemeiner Intelligenz und akkuraten Persönlichkeitsbeurteilungen um eine weitere Studie, die von Colvin und Bundick (2001) beschrieben und in keiner der zuvor genannten Veröffentlichungen berücksichtigt wurde. Unter Verwendung des zuvor erwähnten 12-minütigen Intelligenztests sowie des Profile Accuracy-Ansatzes konnte in dieser Studie kein Zusammenhang gefunden werden. Dies wurde in der Studie unter Bezugnahme auf das RAM von Funder (1995) als hypothesenkonform interpretiert. Die Autoren der Studie vermuteten, dass bei der Beurteilung fremder Personen – wie es in der Studie der Fall war – eher die valide Erkennung als die valide Nutzung persönlichkeitsrelevanter Hinweise eine wichtige Rolle spielt, Intelligenz aber erst für die Nutzung von Relevanz ist (Colvin & Bundick, 2001).

Insgesamt betrachtet findet sich empirische Evidenz für den Zusammenhang zwischen akkuraten Persönlichkeitsbeurteilungen und Intelligenz. Dieser variiert allerdings in der Höhe und liegt auf manifester Ebene maximal im kleinen bis mittleren Bereich. Ein möglicher Grund für die Varianz in den Ergebnissen könnte in der unterschiedlichen Methodik der Studien liegen. Wie bereits im Zusammenhang mit den Ergebnissen berichtet, unterscheiden sich die zitierten Studien dahingehend, ob die akkuraten Beurteilungen entsprechend dem Profile Accuracy-Ansatz oder Trait Accuracy-Ansatz bestimmt wurden. Darüber hinaus wurden in den Studien unterschiedliche Persönlichkeitseigenschaften eingeschätzt (z.B. Extraversion, Neurotizismus und Maskulinität-Femininität bei Lippa und Dietz, 2000; alle Big Five-Faktoren bei Christiansen et al., 2005) sowie unterschiedliche Kriterien verwendet, um das Ausmaß der

Akkuratheit zu bewerten (z.B. Kombination aus Selbstbericht, Fremdbericht durch Bekannte sowie Fremdbericht durch klinische Psycholog:innen bei Letzring, 2008; Fremdberichte durch mehrere Arbeitspsycholog:innen bei De Kock et al., 2015). Diese sowie weitere Unterschiede zwischen den einzelnen Studien könnten die Ergebnisse mit beeinflusst haben.

Der Zusammenhang mit allgemeiner Intelligenz und insbesondere logischem Schlussfolgern wurde darüber hinaus auch für die mit PU verwandten Konstrukte SU und DI betrachtet. SU zeigte in Studien, in denen das Konstrukt entsprechend dem Integrativen Modell Sozialer Intelligenz konzeptualisiert wurde, sowohl auf manifester als auch latenter Ebene maximal geringe Zusammenhänge zu logischem Schlussfolgern. Weis und Süß (2007) berichten für den Zusammenhang zwischen verschiedenen (Unter-)Tests zur Erfassung von SU und logischem Schlussfolgern manifeste Korrelationen zwischen r = -.07 und .19 (alle n.s.). Auch Baumgarten (2015) berichtet ähnliche geringe manifeste Korrelationen zwischen SU-Faktorwerten und logischem Schlussfolgern, die je nach Auswertungsmethode bei r = -.12 bis .12 (alle n.s.) lagen. Latente Zusammenhänge zwischen den beiden Konstrukten sind bei Conzelmann et al. (2013; vgl. auch Seidel, 2007; Weis, 2008) zu finden. Hier ergaben sich je nach Modell ebenfalls geringe latente Zusammenhänge zwischen -.11 (n.s.) bis .20 (Conzelmann et al., 2013). In allen Studien wurde dieser fehlende Zusammenhang als Evidenz für Konstruktvalidität in Form von diskriminanter Validität gedeutet, was im Kontrast zur Konzeptualisierung von SU bei insbesondere Weis (2008) steht. Der Zusammenhang von DI zu allgemeiner Intelligenz und logischem Schlussfolgern fiel hingegen deutlich höher aus. Christiansen et al. (2005) berichten eine manifeste Korrelation von r = .43, p < .01 zwischen DI und allgemeiner Intelligenz und bei De Kock et al. (2015) fiel dieser Zusammenhang mit r = .68, p < . 01 deskriptiv noch größer aus. Die drei Komponenten der DI zeigten mit r = .47 bis .61, alle p < .01, nur leicht geringere Korrelationen zur allgemeinen Intelligenz als DI im Gesamten (De Kock et al., 2015). Manifeste Zusammenhänge zwischen DI und logischem Schlussfolgern berichten de Vries et al. (2021). Für DI im Gesamten ergab sich in dieser Studie ein Zusammenhang von r = .41, p < .001 und für die Komponenten der DI Zusammenhänge von r = .26 bis .37, p < .001.

Der Zusammenhang mit Intelligenz beziehungsweise logischem Schlussfolgern fällt bei den beiden konzeptuell mit PU verwandten Konstrukte SU und DI somit noch einmal deutlich variabler aus als im Bereich akkurater Persönlichkeitsbeurteilungen. In beiden Fällen ist eine Übertragung auf das Konzept PU erneut nur mit Einschränkungen und unter Vorbehalt möglich. So ist SU inhaltlich deutlich breiter konzeptualisiert als PU (Weis, 2008; Weis & Süß, 2005) und Validitätsevidenz der eingesetzten Verfahren zur Erfassung von SU, wie dem MTSI,

ist nur eingeschränkt verfügbar (Conzelmann et al., 2013; Baumgarten, 2015). Bei Weis und Süß (2007) wiesen drei der vier verwendeten Aufgaben zur Erfassung von SU mit einem Cronbachs α von .32 bis .63 zudem sehr geringe Reliabilitäten auf. Die vierte Aufgabe wies ein α von .77 auf und zeigte mit r = .19 den größten Zusammenhang mit logischem Schlussfolgern. Die Ergebnisse der DI lassen sich auf Grund der anderen angenommenen kognitiven Operation ebenfalls nur mit Einschränkungen auf PU übertragen. Wenn man allerdings bedenkt, dass das Wissenskonstrukt DI bereits einen Zusammenhang von r = .41 mit logischem Schlussfolgern aufweist, so kann für PU, für das dieselbe inhaltliche Verankerung wie DI angenommen wird sowie als kognitive Operation das logische Schlussfolgern an sich, ein noch höherer Zusammenhang erwartet werden.

3.1.3 PU und Persönlichkeit

Vor dem Hintergrund der Frage, welche Personen gute Beurteiler:innen der Persönlichkeit anderer Personen sind, wird oftmals auch die Persönlichkeit der Beurteiler:innen betrachtet. Zwar konnten in mehreren Studien Zusammenhänge zwischen akkuraten Persönlichkeitsbeurteilungen und einzelnen Big Five-Faktoren gefunden werden (z.B. Hall et al., 2016; Letzring, 2008; Lippa & Dietz, 2000), allerdings kamen Autor:innen, die eine Übersicht über die Studienlage erstellt haben, zu dem Schluss, dass es keine große Konsistenz in den Ergebnissen gibt (Colman, 2021; De Kock et al., 2020). De Kock et al. (2020) konnten in ihrem Review – in dem nicht nur akkurate Persönlichkeitsbeurteilungen berücksichtigt wurden – lediglich für Offenheit für Erfahrungen ($\bar{r}$ = .10) und Verträglichkeit ($\bar{r}$ = .09) geringe signifikante Zusammenhänge zu akkuraten Beurteilungen finden. Wie Letzring (2008) anmerkte, könnte ein Grund für die Inkonsistenzen in den Ergebnissen auch hier in der oftmals unterschiedlichen Bestimmung der Akkuratheit liegen.

Aus theoretischer Sicht lassen sich vor allem Zusammenhänge zwischen akkuraten Persönlichkeitsbeurteilungen und den beiden für soziale Interaktionen relevanten Big Five-Faktoren Extraversion und Verträglichkeit (McCrae & Costa, 1989) vermuten. Definitionen und Beschreibungen beider Faktoren (z.B. John & Srivastava, 1999; Ostendorf & Angleitner, 2004; vgl. Abschnitt 3.1.1) verdeutlichen deren sozialen Fokus und deuten darauf hin, dass Personen mit einer hohen Ausprägung tendenziell mehr Kontakt mit anderen Menschen haben. Daraus lässt sich wiederum die Annahme ableiten, dass diese Personen mehr Wissen über Persönlichkeitseigenschaften anderer Personen sowie mehr Erfahrung in der Beurteilung dieser Eigenschaften sammeln und folglich auch akkuratere Beurteilungen vornehmen können.

Allerdings lässt sich hieraus nicht zwingend ein Zusammenhang der beiden Faktoren zu PU ableiten, was im Folgenden näher erläutert wird.

Mehr Kontakt mit anderen Personen sollte in zweierlei Hinsicht zu mehr Wissen über Persönlichkeitseigenschaften führen: 1.) Zum einen sollte mehr Kontakt zu mehr individuellem Wissen über die Persönlichkeit einzelner Zielpersonen führen. Dies resultiert dem RAM (Funder, 1995) zufolge in akkurateren Beurteilungen, allerdings unabhängig von der Fähigkeit der Beurteiler:innen. So sind Unterschiede in der Quantität der persönlichkeitsrelevanten Information dem zweiten Schritt des RAM (Verfügbarkeit) zuzuordnen und nicht dem für PU relevanten letzten Schritt (Nutzung). 2.) Zum anderen kann vermutet werden, dass mehr Kontakt zu mehr allgemeinem Wissen über Persönlichkeitseigenschaften führt, beispielsweise in Form von Wissen darüber, welche Persönlichkeitseigenschaften in der Regel gemeinsam bei einer Person auftreten (vgl. Christiansen et al., 2005). Für PU wird allerdings die kognitive Operation logisches Schlussfolgern angenommen, sodass die Aneignung von mehr allgemeinem Wissen nicht zu einer besseren Leistung bei PU führen sollte.

Die Annahme, dass verträgliche und extravertierte Personen mehr allgemeines Wissen über Persönlichkeit erwerben, kann durch die Studien zur DI zudem nicht durchweg bestätigt werden. Bei Christiansen et al. (2005) zeigte DI lediglich einen substantiellen Zusammenhang mit Offenheit für Erfahrungen (r = .34, p < .01) und bei de Vries et al. (2021) zeigten sich maximal kleine Zusammenhänge mit allen Persönlichkeitsfaktoren des HEXACO-Modells. Nur bei De Kock et al. (2015) ergab sich ein substantieller positiver Zusammenhang zwischen DI und Verträglichkeit (r = .25, p < .01). Allerdings fanden die Autor:innen ebenfalls Zusammenhänge zur Gewissenhaftigkeit (r = .20, p < .05) sowie zur Extraversion (r = -.31, p < .01; De Kock et al., 2015). Letzterer steht dabei in Einklang mit den Ergebnissen von Ambady et al. (1995), bei denen sich negative Zusammenhänge zwischen Geselligkeit – einer Komponente der Extraversion (Watson & Clark, 1997) – und der akkuraten Beurteilung von vier Persönlichkeitsdimensionen ergaben (r = -.37, p < .01 bis r = -.23, p < .05). Das lässt die Vermutung zu, dass extravertierte Beurteiler:innen zwar viel Kontakt mit anderen Personen haben, allerdings in den Interaktionen nicht auf die Persönlichkeitseigenschaften der anderen Personen achten oder sich nicht hierfür interessieren und so kein persönlichkeitsrelevantes allgemeines Wissen aneignen.

Der positive Zusammenhang zwischen DI und Verträglichkeit bei De Kock et al. (2015) ist wiederum konsistent mit den Ergebnissen von Biesanz (2010). Dieser konnte im Rahmen seiner Studien zum SAM einen positiven Zusammenhang von Verträglichkeit mit der Perceiver Normative Accuracy (d.h. der Akkuratheit in der Einschätzung der durchschnittlichen

Ausprägungen von Persönlichkeitseigenschaften) finden, nicht aber mit der Perceiver Distinctive Accuracy (d.h. der Akkuratheit in der Einschätzung der individuellen Eigenschaften der Zielpersonen). Ähnliche Ergebnisse resultierten in der Studie allerdings auch für Extraversion, Gewissenhaftigkeit sowie Neurotizismus, wobei sich bei Letzterem ein negativer Zusammenhang zeigte. Für Offenheit für Erfahrungen ergab sich hingegen ein umgekehrtes Muster, da dieser Faktor nur signifikante positive Zusammenhänge zur Perceiver Distinctive Accuracy aufwies (Biesanz, 2010). Letzring (2015) fand für die Verträglichkeit dasselbe Zusammenhangsmuster wie Biesanz, allerdings nicht für die anderen Big Five-Faktoren. Positive Zusammenhänge zwischen akkuraten Persönlichkeitsbeurteilungen unter Verwendung von Profilkorrelationen, die eine normative Komponente der Akkuratheit beinhalten (Funder, 1999; Furr, 2008; Kenny & Winquist, 2001), ergaben sich außerdem bei Letzring (2008) sowie Vogt und Colvin (2003). Im Falle einer direkten Interaktion zwischen Beurteiler:in und Zielperson ist zudem eine alternative Erklärung für den Zusammenhang zwischen Akkuratheit und Verträglichkeit möglich. So vermutete Letzring (2008), dass eine hohe Verträglichkeit von Beurteiler:innen die Wahrscheinlichkeit erhöht, von den Zielpersonen persönlichkeitsrelevante Hinweise zu erhalten. Dies könnte der Autorin zufolge daran liegen, dass sich die Zielpersonen bei verträglichen Beurteiler:innen wohler fühlen und daher eher Information über sich preisgeben. Dies würde erneut den zweiten Schritt des RAM (Verfügbarkeit) betreffen (Letzring, 2008) und nicht den für PU relevanten letzten Schritt.

Auch mehr Erfahrung in der Beurteilung von Persönlichkeitseigenschaften durch vermehrten Kontakt mit anderen Personen sollte keinen großen Einfluss auf PU haben. So hat sich gezeigt, dass die Trainierbarkeit von Gf, die nach dem CHC-Modell das induktive und deduktive logische Schlussfolgern umfasst (McGrew, 2009), wenn überhaupt nur sehr eingeschränkt möglich ist (z.B. Au et al., 2015; Melby-Lervåg & Hulme, 2013; Watrin et al., 2022). Dies kann somit auch für PU angenommen werden.

Insgesamt betrachtet sind für PU auf Grund der Annahme, dass es sich hierbei um eine Form des logischen Schlussfolgerns handelt, ähnliche Zusammenhänge mit Persönlichkeit erwartbar, wie es bei klassischen Formen des logischen Schlussfolgerns der Fall ist. Ackerman und Heggestad (1997) betrachteten in einer Meta-Analyse die Zusammenhänge zwischen verschiedenen intellektuellen Fähigkeiten und Persönlichkeitseigenschaften. Für Gf und die Big Five-Faktoren zeigten sich hierbei Zusammenhänge von -.08 bis .08 (nur zwei Zusammenhänge $p < .05$) und somit keinerlei substantielle Korrelationen. Erwähnenswert sind hier lediglich Zusammenhänge von .30 und. .33 (beide $p < .05$) zwischen Offenheit für Erfahrungen und Gc beziehungsweise allgemeiner Intelligenz (Ackerman & Heggestad, 1997).

Unter Bezugnahme auf diese Meta-Analyse sind für PU und die Big Five-Faktoren somit Zusammenhänge nahe Null zu erwarten. Für das konzeptuell mit PU verwandte Konstrukt SU konnte eine Unabhängigkeit von den Big Five-Faktoren empirisch bereits überwiegend aufgezeigt werden (Conzelmann et al., 2013). In den beiden von Conzelmann et al. (2013) durchgeführten Studien zeigten einzelne SU-Aufgaben zwar Zusammenhänge > |.20| zu Extraversion, Verträglichkeit und Offenheit für Erfahrungen, allerdings ergab sich kein systematisches Korrelationsmuster.

3.1.3.1 Ähnlichkeit zwischen Beurteiler:in und Zielperson. Beurteiler:innen tendieren dazu, die Persönlichkeit der Zielpersonen ähnlich zur eigenen Persönlichkeit einzuschätzen, was Cronbach (1955) sowie Gage und Cronbach (1955) als *angenommene Ähnlichkeit* (assumed similarity) bezeichneten. In dem Zusammenhang thematisiert wird in der Literatur auch die *tatsächliche Ähnlichkeit* (real similarity, actual similarity), also die Übereinstimmung der Persönlichkeitsprofile von Beurteiler:in und Zielperson (Gage & Cronbach, 1955; Paunonen & Hong, 2013). Die angenommene Ähnlichkeit ist deswegen von Interesse, da sie einen möglichen Einflussfaktor auf die Akkuratheit von Persönlichkeitsbeurteilungen darstellt. Sie kann dann zu höherer Akkuratheit führen, wenn beide Personen auch tatsächlich Ähnlichkeit im Hinblick auf die beurteilte Persönlichkeitseigenschaft aufweisen (Gage & Cronbach, 1955; Paunonen & Hong, 2013; Vogt & Colvin, 2003). Dies zeigte sich in einer Studie von Vogt und Colvin (2003), die für den Zusammenhang zwischen *angenommener* Ähnlichkeit und Akkuratheit – beides bestimmt mittels Profilkorrelationen auf individueller Ebene[3] – einen großen Effekt von $r = .66$, $p < .001$ ermittelten. Allerdings variierte dieser Zusammenhang deutlich, wenn die Autoren die *tatsächliche* Ähnlichkeit mitbetrachteten: Drei von vier untersuchten Zielpersonen wiesen im Mittel eine hohe tatsächliche Ähnlichkeit zu den Beurteiler:innen auf – hier zeigte sich weiterhin eine hohe Korrelation zwischen angenommener Ähnlichkeit und Akkuratheit von $r = .64$ bis $.71$ ($p < .001$). Bei der vierten Zielperson, die im Mittel keine tatsächliche Ähnlichkeit zu den Beurteiler:innen aufwies, ergab sich hingegen eine deutlich geringere Korrelation von $r = -.11$, $p < .05$ (Vogt & Colvin, 2003). Dieses Ergebnis macht deutlich, dass bei der Interpretation des Zusammenhangs zwischen angenommener Ähnlichkeit und Akkuratheit die tatsächliche Ähnlichkeit berücksichtigt werden sollte. So ist davon auszugehen, dass angenommene Ähnlichkeit auch zu geringerer Akkuratheit führen kann, und

[3] Die angenommene Ähnlichkeit wurde bestimmt, indem die Selbsteinschätzung je Beurteiler:in mit deren oder dessen Fremdeinschätzung je Zielperson korreliert wurde (Vogt & Colvin, 2003).

zwar wenn keine tatsächliche Ähnlichkeit vorliegt (Paunonen & Hong, 2013). Vogt und Colvin (2003) schlossen aus den Ergebnissen ihrer Studie, dass die angenommene Ähnlichkeit, ebenso wie die Stereotype Accuracy (vgl. Abschnitt 2.3.1), einen validen Prozess im Rahmen der Persönlichkeitsbeurteilung darstellen und vorhandene Hinweise auf die Persönlichkeit der Zielperson sinnvoll ergänzen kann. Die beschriebenen sowie weiterführende Analysen von Vogt und Colvin (2003) deuten den Autoren zufolge darauf hin, dass akkurate Beurteiler:innen möglicherweise nur dann Information über die eigene Persönlichkeit verwenden, wenn sie erkennen, dass die Zielperson tatsächliche Ähnlichkeit zu ihnen aufweist.

Bei Funder et al. (1995) ergaben sich zwischen *tatsächlicher* Ähnlichkeit und Akkuratheit mittlere Korrelationen, wenn fremde Zielpersonen beurteilt wurden *(r* = .33 bzw. .40, $p < .001$) sowie große Effekte, wenn die Zielpersonen bekannt waren (r = .51 bzw. .55, $p < .001$). Sowohl die Akkuratheit als auch die Ähnlichkeit wurden in dieser Studie ebenfalls auf individueller Ebene mittels Profilkorrelationen bestimmt.[4] Funder et al. (1995) führten darüber hinaus multiple Regressionen durch, in denen die Selbsteinschätzungen der Zielpersonen sowohl durch die Fremd- als auch Selbsteinschätzungen der Beurteiler:innen vorhergesagt wurden. Bei bekannten Zielpersonen wiesen die Fremdeinschätzungen der Beurteiler:innen im Mittel ein größeres Regressionsgewicht auf als die Selbsteinschätzungen der Beurteiler:innen – bei fremden Zielpersonen ergab sich hingegen das umgekehrte Muster. Die Autor:innen schlossen aus diesen Ergebnissen, dass bei der Beurteilung fremder Zielpersonen eher auf Information über ihre eigene Persönlichkeit zurückgegriffen wird, da hier ausreichende Hinweise auf die Persönlichkeit der Zielpersonen fehlen (Funder et al., 1995).

Eine mögliche Erklärung für den Zusammenhang zwischen tatsächlicher Ähnlichkeit und Akkuratheit findet sich im RAM (Funder, 1995). Wie in Abschnitt 2.2.2 erwähnt, schlug Funder (1995) mögliche Einflüsse von zweifach-Interaktionen der im Rahmen des RAM diskutierten Moderatoren auf die Akkuratheit vor. Die Interaktion zwischen den Moderatoren Good Judge und Good Trait, die sogenannte *Expertise*, beschreibt die Annahme, dass bestimmte Beurteiler:innen bestimmte Persönlichkeitseigenschaften besser einschätzen können (Funder, 1995). So ist es laut Funder denkbar, dass auf Grund von Erfahrung oder explizitem Lernen unterschiedlich viel Wissen über eine bestimmte Persönlichkeitseigenschaft vorhanden ist. Dieses Wissen könnte wiederum dazu beitragen, dass Information über diese

[4] Die tatsächliche Ähnlichkeit wurde bestimmt, indem die Selbsteinschätzung je Beurteiler:in mit der Selbsteinschätzung je Zielperson korreliert wurde (Funder et al., 1995).

Eigenschaft akkurater wahrgenommen und vor allem genutzt wird (Funder, 1995). Darauf aufbauend lässt sich vermuten, dass Beurteiler:innen Expert:innen hinsichtlich ihrer eigenen Persönlichkeitsausprägungen sind und diese bei anderen Personen besser beurteilen können (Letzring & Funder, 2021). Gegen diese Erklärung sprechen allerdings die Ergebnisse von Kurtz und Sherker (2003) sowie Hartung und Renner (2011), die die tatsächliche Ähnlichkeit und Akkuratheit je Big Five-Faktor betrachtet sowie mittels Regressionsanalysen potentielle Moderatoreffekte der tatsächlichen Ähnlichkeit überprüft haben. Kurtz und Sherker (2003) fanden nur für Offenheit für Erfahrungen und Neurotizismus einen signifikanten Moderatoreffekt. Bei diesen beiden Faktoren erzielten Beurteiler:innen eine noch höhere Akkuratheit, wenn Beurteiler:in und Zielperson tatsächlich ähnliche Ausprägungen aufwiesen (Kurtz & Sherker, 2003). Allerdings zeigten die Interaktionen jeweils nur eine geringe inkrementelle Varianzaufklärung von rund 3 %. Hartung und Renner (2011) konnten auf Basis ähnlicher Analysen für keinen der Big Five-Faktoren einen signifikanten Interaktionseffekt finden. Nach Letzring und Funder (2021) spricht dieses Ergebnis – und somit auch das von Kurtz und Sherker (2003) – gegen die Annahme der Expertise-Interaktion von Funder (1995).

Im Gegensatz zu den zuvor berichteten Ergebnissen zeigte sich bei Letzring (2015) auf Basis von Moderatoranalysen im SAM von Biesanz (2010) ein Zusammenhang zwischen der Persönlichkeit der Beurteiler:innen und deren Akkuratheit bei der Beurteilung der Zielpersonen, wenn jeweils derselbe Big Five-Faktor betrachtet wurde. Dieser Zusammenhang fiel bei Extraversion, Verträglichkeit, Gewissenhaftigkeit und Offenheit positiv aus, bei Neurotizismus hingehen negativ (Letzring, 2015). Allerdings konnte Letzring (2015) dieses Ergebnis nur für die Normative Accuracy und nicht für die Distinctive Accuracy nachweisen. Ähnliche Ergebnisse ergaben sich bei Human und Biesanz (2011, 2012). Hier zeigten sich keine signifikanten Zusammenhänge zwischen einem Maß für *angenommene* Ähnlichkeit, bei dem unter anderem die tatsächliche Ähnlichkeit kontrolliert wurde, und der Distinctive Accuracy (r = -.08 und .17 bei Human & Biesanz, 2011; r = -.13 bei Human & Biesanz, 2012), aber zum Teil signifikante Zusammenhänge zwischen dem Maß für angenommene Ähnlichkeit und der Normative Accuracy (Human & Biesanz, 2011; Studie 1: r = .33, p < .001; Studie 2: r = .11, n.s.). Diese Ergebnisse stehen zudem nicht im Kontrast zu den Ergebnissen von Vogt und Colvin (2003) und Funder et al. (1995), die Profilkorrelationen verwendeten, bei denen eine normative Komponente der Akkuratheit enthalten ist (Funder, 1999; Winquist & Kenny, 2001). Allerdings sprechen die Ergebnisse von Letzring (2015) sowie Human und Biesanz (2011, 2012) ebenfalls gegen die Annahme, dass Beurteiler:innen bei anderen Personen insbesondere die eigenen Persönlichkeitsausprägungen gut beurteilen können (Funder, 1995).

In einigen Studien wurden die Akkuratheits- und Ähnlichkeitswerte nicht auf individueller Ebene je Beurteiler:in, sondern für einzelne Persönlichkeitsdimensionen über alle Beurteiler:innen hinweg betrachtet. Beispielsweise bestimmten Watson et al. (2000) für die Big Five sowie verschiedene Affektskalen sowohl Akkuratheitswerte als auch Werte für die *angenommene* Ähnlichkeit. Hierbei zeigte sich, dass bei weniger akkurat beurteilten Dimensionen tendenziell eine höhere angenommene Ähnlichkeit vorlag (r = -.77, p < .01 bis r = -.29, n.s.; für ähnliche Ergebnisse siehe Beer & Watson, 2008; Human & Biesanz, 2012; Paunonen & Hong, 2003; siehe aber auch Lee et al., 2009). Paunonen und Hong (2013) konnten zudem auf Basis eigener Analysen und (Re-)Analysen anderer Studien aufzeigen, dass die Akkuratheitswerte tendenziell für diejenigen Skalen am geringsten ausfallen, bei denen die größte Diskrepanz zwischen angenommener und tatsächlicher Ähnlichkeit besteht. Ready et al. (2000) bestimmten neben der Akkuratheit sowohl die tatsächliche als auch die angenommene Ähnlichkeit. Zum einen zeigte sich auf Ebene der Persönlichkeitseigenschaften erneut eine negative Korrelation zwischen Akkuratheit und *angenommener* Ähnlichkeit (r = -.31, n.s.). Im Kontrast dazu zeigte sich eine positive Korrelation (r = .26, n.s.) zwischen Akkuratheit und *tatsächlicher* Ähnlichkeit. Darüber hinaus konnten Ready et al. (2000) zeigen, dass tendenziell für solche Persönlichkeitseigenschaften eine höhere Ähnlichkeit angenommen wird, die als bei anderen schwierig einschätzbar wahrgenommen werden (r = -.73, p < .05).

Laut Watson et al. (2000) stützt der negative Zusammenhang zwischen Akkuratheit und angenommener Ähnlichkeit auf Ebene der Persönlichkeitsdimensionen die Annahme von Funder et al. (1995), dass Information über die eigene Persönlichkeit dann verwendet wird, wenn die vorhandene Information über die Persönlichkeit der Zielperson nicht ausreicht (vgl. auch Ready et al., 2000). Paunonen und Hong (2013) gehen ebenfalls davon aus, dass Personen immer dann auf Information über die eigene Persönlichkeit zurückgreifen, wenn keine valide Information über die Zielperson vorhanden ist. Allerdings betonen die Autoren, dass diese Strategie vermutlich den kleinsten Varianzanteil in den Beurteilungen aufklären kann und erst dann genutzt wird, wenn keine direkte behaviorale Information über die einzuschätzende Persönlichkeitseigenschaft, keine andere behaviorale Information über die Zielperson sowie keine Information, die die Anwendung von allgemeinem Wissen über Stereotypen erlaubt, vorliegt (Paunonen & Hong, 2013).

Die Ergebnisse von Lee et al. (2009) sprechen nach eigener Aussage wiederum gegen die zuvor beschriebene Erklärung der angenommenen Ähnlichkeit. In ihrer Studie konnten die Autor:innen für die zwei HEXACO-Faktoren Ehrlichkeit-Bescheidenheit und Offenheit für Erfahrungen sowohl hohe Akkuratheitswerte (vergleichbar mit denen der anderen Faktoren)

als auch hohe Werte für die angenommene und tatsächliche Ähnlichkeit finden (höher als für die anderen Faktoren). Lee et al. (2009) erklärten die hohen Ähnlichkeitswerte für die beiden akkurat beurteilten Faktoren mit deren Relevanz für das persönliche Wertesystem, das eine wichtige Rolle in sozialen Beziehungen spiele. Thielmann et al. (2020, 2022) konnten auf Basis mehrerer Studien sowohl die Ergebnisse als auch die Schlussfolgerung von Lee et al. (2009) stützen. Den Autor:innen zufolge ist die Relevanz für die persönlichen Werte derjenige Faktor, der beeinflusst, welche Persönlichkeitseigenschaften bei Fremden als ähnlich zur eigenen Persönlichkeit angenommen werden. Darüber hinaus fanden Thielmann et al. (2020, 2022) keine Unterstützung für die Annahme, dass Information über die eigene Persönlichkeit dann verwendet wird, wenn die vorhandene Information über die Zielperson nicht ausreicht.

Insgesamt betrachtet ist die Evidenz für einen möglichen Effekt von Ähnlichkeit zwischen Beurteiler:in und Zielperson auf die Akkuratheit von Persönlichkeitsbeurteilungen sehr gemischt. Dies liegt vermutlich nicht zuletzt an den auch hier zum Teil sehr heterogenen methodischen Vorgehensweisen der Autor:innen, was das Zusammenfassen der Ergebnisse deutlich erschwert. Nichtsdestotrotz deutet ein Teil der Ergebnisse darauf hin, dass angenommene Ähnlichkeit dann einen positiven Zusammenhang mit Akkuratheit aufweist, wenn tatsächliche Ähnlichkeit vorliegt (Funder et al., 1995; Paunonen & Hong, 2013; Vogt & Colvin, 2003). Einschränkend muss allerdings beachtet werden, dass dieser Zusammenhang möglicherweise nur für die normative Komponente der Akkuratheit vorhanden ist (vgl. Human & Biesanz, 2011, 2012; Letzring, 2015). Eine theoretische Erklärung für den möglichen Effekt der Ähnlichkeit auf die Akkuratheit scheint noch nicht abschließend geklärt. So werden immer noch verschiedene Gründe, warum Beurteiler:innen andere Personen ähnlich zu sich selbst einschätzen, diskutiert und überprüft (vgl. Thielmann et al., 2020, 2022). Dadurch ist die Relevanz der angenommenen und tatsächlichen Ähnlichkeit bei PU ebenfalls nicht eindeutig abschätzbar. Sollte sich der Effekt vor allem auf die normative Komponente der Akkuratheit (d.h. eine Wissenskomponente; vgl. Biesanz, 2010) beziehen, so kann aus theoretischer Sicht eine Abgrenzung zu PU vorgenommen werden. Handelt es sich, wie von Vogt und Colvin (2003) vermutet, um einen validen Prozess im Rahmen der Persönlichkeitsbeurteilung, der die vorhandenen Hinweise sinnvoll ergänzen kann, ist auch bei PU ein Zusammenhang zur Ähnlichkeit nicht auszuschließen und sollte daher zumindest bei der Operationalisierung der Fähigkeit als möglicher Einflussfaktor mitbedacht werden.

3.2. Operationalisierung

Abschließend wird die Frage nach einer angemessenen Operationalisierung von PU thematisiert. Entsprechend der Konzeptualisierung als Fähigkeit zum logischen Schlussfolgern im Bereich Persönlichkeit handelt es sich bei PU um ein Fähigkeitskonstrukt. Folglich erfordert die Erfassung den Einsatz eines Leistungstests, bei dem die Testpersonen Schlussfolgerungen über Persönlichkeitseigenschaften anderer Personen ziehen müssen, die im Anschluss eindeutig als richtig oder falsch bewertet werden.

In der Literatur finden sich eine Reihe an verschiedenen Ansätzen zur Bestimmung und Analyse der Akkuratheit von Persönlichkeitsbeurteilungen, auf die man prinzipiell bei der Erfassung von PU zurückgreifen könnte. Wie sich allerdings in Abschnitt 2.3 gezeigt hat, sind diese Ansätze zum Teil sehr heterogen. So müsste unter anderem eine Wahl zwischen dem Ansatz der Profile Accuracy und dem der Trait Accuracy getroffen werden, die nicht trivial erscheint (vgl. Hall et al., 2018). Bei Verwendung des Profile Accuracy-Ansatzes müsste darüber hinaus eine Entscheidung für oder gegen die Kontrolle der enthaltenen Stereotype Accuracy-Komponente getroffen werden, wobei Argumente für beide Richtungen vorliegen (z.B. Colvin & Bundick, 2001; Funder, 1999). Diese Heterogenität ist allerdings nicht die zentrale Problematik, sondern eine noch viel grundlegendere Herausforderung, von der alle bisher verfügbaren Ansätze betroffen sind: Um bei einem Leistungstest zur Erfassung von PU die Schlussfolgerungen der Testpersonen eindeutig als richtig oder falsch bewerten zu können, wird eine korrekte Antwort, das heißt die wahre Ausprägung auf der oder den latenten Persönlichkeitseigenschaft(en) der Zielperson benötigt.

Die Problematik hinsichtlich der Bestimmung der wahren Persönlichkeitsausprägung der Zielpersonen wurde bereits früh (Vernon, 1933) und seither von verschiedenen Autor:innen diskutiert (z.B. Bernieri, 2001; Funder, 1999; Kenny, 1994). In diesem Punkt besteht eine weitere Parallele zum Forschungsbereich der EI, wo ebenfalls Schwierigkeiten in der Bestimmung einer eindeutig korrekten Antwort bestehen. Während in klassischen Intelligenztestaufgaben formale Systeme oder Regeln wie Mathematik oder Logik verwendet werden, um objektiv richtige Antworten zu identifizieren, fehlt im Inhaltsbereich Emotionen ein solches Rational (Hellwig et al., 2020; MacCann et al., 2004; Roberts et al., 2001). Diese Argumentation ist auf den Inhaltsbereich Persönlichkeit direkt übertragbar. Daher ist es auch nicht verwunderlich, dass in beiden Bereichen ähnliche bis identische alternative Kriterien verwendet werden, um die wahre Ausprägung und somit eine korrekte Antwort festzulegen (vgl. Funder, 2012; Kenny, 1994; Mayer et al., 1999; Roberts et al., 2001). Wie bereits bei der

Vorstellung der Kriterien aus dem Bereich akkurater Persönlichkeitsbeurteilungen in Abschnitt 2.3 erwähnt wurde, sind diese alternativen Kriterien allerdings problembehaftet und erlauben keine eindeutige Bestimmung einer korrekten Antwort und somit keine Konstruktion eines klassischen Leistungstests. Auch hier wird in beiden Forschungsbereichen auf sehr ähnliche bis identische Probleme hingewiesen, die im Folgenden erläutert werden.

Mögliche Probleme bei der Verwendung des Selbstberichts der Zielperson als Kriterium – was im Bereich EI auch als Target Scoring bezeichnet wird (Mayer et al., 1999) – bestehen darin, dass die Zielperson möglicherweise nicht zu einer validen Auskunft fähig ist oder verzerrte Antworten gibt (z.B. sozial erwünschte Antworten; Bernieri, 2001; Funder, 2012; Kenny, 1994; Mayer & Geher, 1996). Interessanterweise ist der Selbstbericht im Bereich akkurater Persönlichkeitsbeurteilungen trotzdem das am häufigsten verwendete Kriterium (Funder, 2012), während es im Bereich EI eher selten Verwendung findet (Roberts et al., 2001).

Wird der Konsensus einer größeren Gruppe an Personen als Kriterium verwendet (Consensus Scoring im Bereich EI; Mayer et al., 1999), kann es vorkommen, dass trotz vorhandener großer Übereinstimmung alle Gruppenmitglieder falsch liegen (Funder, 2012; Kenny, 1994; Roberts et al., 2001). Im Bereich EI werden zudem statistische und methodische Probleme dieser Scoring Variante diskutiert (MacCann et al., 2004).

Die Verwendung der Urteile von Expert:innen als Kriterium (Expert Scoring im Bereich EI; Mayer et al., 1999) bringt vor allem die Frage mit sich, wer Expert:in ist und zudem ist es denkbar, dass sich Expert:innen widersprechen oder ebenfalls falsch liegen (Bernieri, 2001; Kenny, 1994; MacCann et al., 2004; Roberts et al., 2001).

Im Bereich EI findet sich zudem das Theorie-basierte Scoring, bei dem die korrekte Antwort auf Basis von Emotionsbewertungstheorien festgelegt wird (Hellwig & Schulze, 2021). Im Bereich akkurater Persönlichkeitsbeurteilungen ist die Verwendung Theorie-basierter Kriterien nicht direkt zu finden. Lediglich beim DIT zur Erfassung von DI sowie beim TOPI zur Erfassung der PI wurden die korrekten Antworten auf Basis konkreter und etablierter Forschungsergebnisse identifiziert (Christiansen et al., 2005; De Kock et al., 2015; Mayer et al., 2012, 2019), was einem Theorie-basierten Scoring nahekommt. Für PU kommt dieses Scoring allerdings nicht in Frage, da für die Bewertung von Schlussfolgerungen über individuelle Zielpersonen keine allgemeinen Persönlichkeitsmodelle verwendet werden können. Darüber hinaus besteht ein Problem des Theorie-basierten Scorings darin, dass dessen Qualität stets von der Richtigkeit der verwendeten Theorie abhängt (Hellwig & Schulze, 2021).

Im Abschnitt 2.3 wurden zudem behaviorale Kriterien beschrieben, die im Bereich EI nicht zum Einsatz kommen. Auch wenn die Verwendung von Verhaltensweisen der Zielperson

als Kriterien zum Teil als Goldstandard bezeichnet wurde (Funder, 1999, 2012), ist auch dieses Kriterium problembehaftet. So muss zunächst einmal ein für die Persönlichkeitseigenschaft relevantes, das heißt korrektes Verhaltenskriterium gefunden und dieses dann auch noch reliabel und valide erfasst werden, was ebenfalls mit Herausforderungen verbunden ist (Bernieri, 2001; Funder, 1999, 2012; Kenny, 1994).

Insgesamt ist somit festzuhalten, dass alle in der Literatur vorgeschlagenen Kriterien zur Bestimmung der wahren Ausprägung der Persönlichkeitseigenschaft einer Zielperson problembehaftet sind. Keines der Kriterien erlaubt ein eindeutiges Scoring der Testpersonenantworten, sodass sich auch keiner der in Abschnitt 2.3 vorgestellten Ansätze für eine angemessene Erfassung einer Fähigkeit wie PU eignet. In diesem Punkt besteht eine weitere Parallele zwischen PU und EU (sowie auch SU; Conzelmann et al., 2013; Weis, 2008), sodass beide Konstrukte nicht nur konzeptuelle Gemeinsamkeiten aufweisen, sondern auch von sehr ähnlichen Herausforderungen bei der Operationalisierung betroffen sind.

3.2.1 Acquisition-Application Testdesign

Beim Acquisition-Application (AcquA) Testdesign von Schulze und Roberts (2015) handelt es sich um einen Ansatz zur Konstruktion von Leistungstests, der diese Scoring-Probleme lösen und eine eindeutige Bewertung der Testpersonenantworten ermöglichen soll. Das Testdesign wurde ursprünglich unter dem Namen Empathic Agent Paradigma (EAP) entwickelt und für die Erfassung von EU vorgeschlagen (Hellwig et al., 2020; Orchard et al., 2009; Schulze et al., 2009), wofür es mittlerweile auch erfolgreich eingesetzt wurde (Hellwig, 2016; Hellwig et al., 2020). Da das Testdesign allerdings nicht ausschließlich für die Erfassung von EU eingesetzt werden kann (Schulze & Roberts, 2015), wird im Folgenden die allgemeinere Bezeichnung AcquA-Testdesign verwendet.

Aufgaben, die nach den Prinzipien des AcquA-Testdesigns konstruiert wurden, bestehen aus zwei Phasen (Hellwig et al., 2020; Schulze & Roberts, 2015; die Quellen dienten als Grundlage für den gesamten Absatz): In der ersten Phase (Aneignungsphase) muss die Testperson Wissen über die Ereignis-Reaktions-Kontingenzen mindestens einer Zielperson erwerben, das heißt darüber, wie die Zielperson typischerweise auf bestimmte Ereignisse reagiert. Die zweite Phase (Anwendungsphase) ist so gestaltet, dass die Testperson ihr zuvor erworbenes Wissen anwenden muss, um auf Basis dessen die Wahrscheinlichkeit verschiedener möglicher Reaktionen der Zielperson in einer neuen, vergleichbaren Situation einzuschätzen. Die Ereignis-Reaktions-Kontingenzen stellen dabei das Kernelement des AcquA-Testdesigns dar. Mit ihnen werden die Regelhaftigkeiten im Verhalten und Erleben der

Zielperson, also beispielsweise, dass sie bei der Konfrontation mit Ereignis X die Reaktion Y zeigt, vorab festgelegt. Entscheidend ist dabei, dass die vorab festgelegten Kontingenzen im Rahmen der Testkonstruktion in die Aneignungsphase eingebaut werden, womit sichergestellt wird, dass die Testperson die Regelhaftigkeiten prinzipiell identifizieren kann. Dies stellt auch die Anforderung an die Testperson in der ersten Phase dar. In der zweiten Phase wird schließlich ein zu Ereignis X aus der Aneignungsphase vergleichbares, aber neues Ereignis X' präsentiert. Zudem werden mehrere mögliche Reaktionen der zuvor kennengelernten Zielperson auf X' präsentiert, die von der Testperson dahingehend beurteilt werden müssen, wie wahrscheinlich oder unwahrscheinlich sie für die Zielperson wären. Hier muss die Testperson die zuvor identifizierte Regelhaftigkeit somit auf ein neues Problem anwenden. Auf Basis der in der Aneignungsphase präsentierten Kontingenz können die möglichen Reaktionen der Zielperson auf X' logikbasiert und damit eindeutig als richtig (Reaktion entspricht Y) oder falsch (Reaktion entspricht nicht Y) bewertet werden.

Bisher wurde das AcquA-Testdesign ausschließlich zur Erfassung von EU eingesetzt (EAP-Test; Hellwig, 2016; Hellwig et al., 2020). Hier besteht die Hauptcharakteristik der AcquA-Aufgaben in der Verwendung von Ereignis-*Emotions*-Kontingenzen, mit denen modelliert wird, welche emotionalen Reaktionen die Zielpersonen in bestimmten Situationen zeigen (Hellwig et al., 2020). In mehreren Studien ergaben sich gute Ergebnisse hinsichtlich der psychometrischen Qualität der neu konstruierten AcquA-EU Aufgaben (Hellwig et al., 2020; Schulze & Jobmann, 2016). Hellwig et al. (2020) konnten durch Untersuchung der Zusammenhänge zu verschiedenen Intelligenzmaßen (logisches Schlussfolgern, Wissen, Merkfähigkeit) sowie einem weiteren neuen Testverfahren zur Erfassung von EU (Hellwig & Schulze, 2021) Evidenz für konvergente Validität sammeln. Zudem lieferte die Untersuchung der Zusammenhänge zu den Big Five-Faktoren Evidenz für diskriminante Validität (Hellwig et al., 2020). Darüber hinaus konnten Hellwig et al. in einer experimentellen Untersuchung die Relevanz der Aneignungsphase und der Kontingenzen für das AcquA-Testdesign aufzeigen. In einer unabhängigen Untersuchung mit neuem Aufgabenmaterial zeigten sich zudem erwartungskonforme Zusammenhänge zu Arbeitsgedächtniskapazität, Merkfähigkeit (erfasst mit Material der AcquA-EU Aufgaben), selbsteingeschätzter Empathiefähigkeit sowie erneut den Big Five-Faktoren (Schulze & Jobmann, 2016). Insgesamt erscheinen die Ergebnisse zur Erfassung von EU mit Hilfe des AcquA-Testdesigns also sehr vielversprechend.

Bedenkt man die hohe konzeptuelle Ähnlichkeit zwischen PU und EU sowie die parallelen Herausforderungen bei der Erfassung der beiden Fähigkeiten, liegt eine Anwendung des AcquA-Testdesigns zur Erfassung von PU nahe. Hellwig (2016) schlug die Anwendung

des Testdesigns zur Erfassung einer Fähigkeit zu akkuraten Persönlichkeitsbeurteilungen sogar explizit vor. Darüber hinaus verglichen Hellwig et al. (2020) die Kontingenzen der AcquA-Aufgaben mit den Wenn-Dann-Profilen aus dem CAPS (Kammrath et al., 2005; Mischel & Shoda, 1995), die auch bei der Ableitung von PU berücksichtigt wurden. Die Vorteile des Testdesigns im Bereich EU, die von Hellwig et al. (2020) beschrieben werden, lassen sich zudem fast vollständig auf den Bereich PU übertragen. Der wichtigste Aspekt ist hierbei, dass das Testdesign die Konstruktion von Aufgabenmaterial ermöglicht, bei dem die korrekte Antwort nicht durch die Verwendung von problembehafteten Kriterien festgelegt werden muss, sondern sich eindeutig aus dem Vorgehen bei der Testkonstruktion ergibt (Hellwig et al., 2020; Schulze & Roberts, 2015). Darüber hinaus können alltagsnahe Situationen simuliert werden, in denen normalerweise Schlussfolgerungen über die Persönlichkeit anderer Personen auf Basis von Information über vergangenes Verhalten gezogen werden (Hellwig et al., 2020). Wie aus den Beschreibungen von Hellwig et al. hervorgeht, ermöglichen die Kontingenzen zudem die Modellierung individueller Persönlichkeitsprofile und damit verschiedener Ausprägungen und Kombinationen von Persönlichkeitseigenschaften.

Rogers und Biesanz (2019) schlussfolgerten auf Basis ihrer Studienergebnisse, dass „for assessing the good judge, it is imperative to focus solely on the context of the good target" (S. 13) sowie dass insbesondere bei der Konstruktion standardisierter Testverfahren entsprechende Zielpersonen systematisch ausgewählt werden müssen. Mit Hilfe des AcquA-Testdesigns kann durch die gezielte Festlegung und Präsentation der Kontingenzen sichergestellt werden, dass alle relevanten und für die vorzunehmenden Schlussfolgerungen benötigten Hinweise für die Testperson verfügbar sind. Es kann damit sichergestellt werden, dass gute Zielpersonen im Sinne der ersten beiden Schritte des RAM (Verfügbarkeit und Relevanz; Funder, 1995) und im Sinne der Forderung von Rogers und Biesanz (2019; vgl. auch Jaksic & Schlegel, 2020) in den Aufgaben verwendet werden. Die Verwendung des AcquA-Testdesigns ähnelt zudem der Verwendung von operationalen Kriterien (vgl. Abschnitt 2.3), bei denen die korrekte Antwort per Definition oder experimenteller Manipulation eindeutig festgelegt wird (Bernieri, 2001; Kenny, 1994). Kenny (1994) ging davon aus, dass operationale Kriterien im Bereich von Persönlichkeitsbeurteilungen kaum anwendbar sind. Durch das AcquA-Testdesign könnte diese Möglichkeit nun geschaffen werden.

3.2.2 Anwendung des AcquA-Testdesigns zur Erfassung von PU

Um das AcquA-Testdesign zur Erfassung von PU anwenden zu können, müssen Anpassungen an den neuen Inhaltsbereich vorgenommen werden. Wie bereits berichtet, sind die Ereignis-

Reaktions-Kontingenzen das Kernelement des Testdesigns, da durch sie festgelegt wird, welche Regelhaftigkeiten in der Aneignungsphase identifiziert werden müssen und damit welches Wissen erworben wird (Schulze & Roberts, 2015). Folglich ist die Auswahl geeigneter Kontingenzen für eine valide Erfassung von PU entscheidend. Sinnvoll erscheint hierbei die Verwendung von Ereignis-*Verhaltens*-Kontingenzen. Persönlichkeitseigenschaften einer Person zeigen sich unter anderem in ihren typischen Verhaltensweisen, sodass aus diesen wiederum Rückschlüsse auf die zugrunde liegende(n) Persönlichkeitseigenschaft(en) gezogen werden können (Cattell, 1979; McCrae & Costa, 2003; Mischel & Shoda, 1995; vgl. Abschnitt 2.1). Darüber hinaus legen Modelle aus dem Bereich akkurater Persönlichkeitsbeurteilungen ebenfalls den Fokus auf das Verhalten der Zielpersonen als Grundlage für die vorzunehmenden Beurteilungen (z.B. RAM; Funder, 1995). Für die Anwendung des AcquA-Testdesigns zur Erfassung von PU werden daher Ereignis-Verhaltens-Kontingenzen genutzt, um die Persönlichkeitseigenschaften der Zielpersonen über deren typische Verhaltensweisen zu modellieren. Prinzipiell wäre auch eine Integration von Emotionen in die Kontingenzen denkbar, da Emotionen einen wichtigen Bestandteil bestimmter Persönlichkeitseigenschaften darstellen (z.B. Neurotizismus und Extraversion; McCrae & John, 1992; Watson & Clark, 1997). Zudem kann ein Überschneidungsbereich zwischen PU und EU angenommen werden (vgl. Abschnitt 3.1.1). In der vorliegenden Arbeit werden Emotionen trotzdem weitestgehend aus den Kontingenzen ausgeschlossen, um neben einer theoretischen auch eine empirische Abgrenzung von PU und EU zu ermöglichen sowie eine mehrdeutige Interpretation der Ergebnisse zu vermeiden.

Wie im Bereich EU ist es auch im Bereich PU durch das AcquA-Testdesign möglich, die Kontingenzen unabhängig von theoretischen Annahmen zu konstruieren (vgl. Hellwig, 2016). Trotzdem erscheint zur Strukturierung und Auswahl der Persönlichkeitseigenschaften, die durch die Kontingenzen modelliert werden sollen, eine Orientierung an den Big Five-Faktoren sinnvoll. Grund hierfür ist zum einen die Entstehung der Big Five. Einer der Ausgangspunkte waren Analysen der natürlichen Sprache zur Beschreibung von Personen (für eine Übersicht siehe Digman, 1990; John et al., 1988). Dieser sogenannte lexikalische Ansatz basiert auf der Annahme, dass über die Zeit für die relevantesten und auffälligsten Unterschiede zwischen Personen Wörter entstanden sind (John et al., 1988). Laut Srivastava (2010) sollten das FFM und die Big Five auf Grund dieser Entstehungsgrundlage in erster Linie als Modell der sozialen Wahrnehmung angesehen werden. Auch Saucier und Goldberg (1996) waren der Ansicht, dass sich die Big Five auf die wahrgenommene Persönlichkeit beziehen. In den Studien, die zur Entstehung der Big Five und des FFM beigetragen haben, wurden zudem

vielfach Fremdberichte der Persönlichkeit verwendet, entweder zusammen mit Selbstberichten oder ausschließlich (z.B. Goldberg, 1990; Norman, 1963; Ostendorf, 1990; Tupes & Christal, 1992), und die Big Five zeigten sich hier sowohl in Selbst- als auch Fremdberichtsdaten (z.B. Goldberg, 1990). Darüber hinaus können die Big Five als integrative deskriptive Taxonomie angesehen werden (John et al., 2008). Die Big Five liefern somit einen Orientierungsrahmen dafür, wie die Persönlichkeit anderer Personen üblicherweise wahrgenommen wird und stellen damit auch einen sinnvollen inhaltlichen Orientierungsrahmen bei der Operationalisierung von PU dar. Nichtsdestotrotz sollte die Operationalisierung nicht von der Korrektheit eines spezifischen Modells abhängig gemacht werden. Bei der Verwendung des AcquA-Testdesigns erfolgen die Schlussfolgerungen daher nicht auf der globalen Ebene der Big Five-Faktoren oder deren Facetten, wo die Leistung der Testpersonen auch von ihrem Wissen hinsichtlich des zugrunde gelegten Modells abhängen würde. Die Schlussfolgerungen werden auf Ebene spezifischer Verhaltensweisen erfasst, einer hierarchisch niedrigeren Ebene unterhalb der Faktoren und Facetten (Digman, 1990). Auf diese Weise besteht keine Abhängigkeit von der Korrektheit des verwendeten Modells. Das FFM und die Big Five helfen somit ausschließlich, die Kontingenzen inhaltlich zu organisieren sowie die gesamte Breite der Persönlichkeit abzudecken. Daher sei betont, dass prinzipiell auch andere Modelle hierfür verwendet werden könnten (z.B. HEXACO-Modell; Ashton & Lee, 2007).

Zusammenfassend handelt es sich beim AcquA-Testdesign von Schulze und Roberts (2015) um einen geeigneten und vielversprechenden Ansatz für die Operationalisierung der Fähigkeit PU. Der entscheidende Vorteil des Testdesigns besteht darin, dass es eines der zentralsten Probleme aus dem Bereich akkurater Persönlichkeitsbeurteilungen adressiert und anders als bisher die eindeutige Identifikation einer korrekten Antwort ermöglicht. Das Ziel des empirischen Teils der vorliegenden Arbeit bestand daher in der Anwendung des AcquA-Testdesigns für die Konstruktion von Leistungstestaufgaben zur Erfassung von PU. Zudem sollte die psychometrische Qualität der neu konstruierten AcquA-PU Aufgaben überprüft und erste Validitätsevidenz gesammelt werden.

4. Studie 1

4.1 Ziele

Das erste Ziel der ersten Studie bestand in der Neukonstruktion von Aufgaben zur Erfassung von PU entsprechend den Prinzipien des AcquA-Testdesigns sowie in der Überprüfung der psychometrischen Qualität. Wie noch näher erläutert wird, sollte auch ein neues Format der Aufgabendarstellung umgesetzt werden. Das erste Ziel beinhaltete zudem die Selektion geeigneter Items sowie eine Überprüfung der auf Grund der Konzeptualisierung von PU angenommenen Eindimensionalität der Aufgaben. Darüber hinaus erfolgte eine Schätzung der Reliabilität, für die auf Grund der Neukonstruktion keine spezifischen Erwartungen formuliert wurden.

Das AcquA-Testdesign wurde bisher ausschließlich für die Konstruktion von Aufgaben zur Erfassung von EU eingesetzt (Hellwig et al., 2020). Auch wenn es prinzipiell für die Erfassung unterschiedlicher Fähigkeitskonstrukte eingesetzt werden kann (Schulze & Roberts, 2015), wurde im Rahmen des zweiten Ziels erstmalig untersucht, ob es neben EU die Erfassung weiterer und insbesondere konzeptuell sehr ähnlicher Konstrukte wie PU erlaubt. Von besonderer Bedeutung war daher die Überprüfung, ob PU und EU empirisch voneinander separiert werden können, wenn beide Konstrukte durch das AcquA-Testdesign operationalisiert werden. Auf Grund der Annahme, dass es sich bei PU und EU um konzeptuell sehr ähnliche, aber dennoch inhaltlich unterscheidbare Konstrukte handelt (vgl. Abschnitt 3.1.1) sowie auf Grund der stark überlappenden Operationalisierung, wurde erwartet, dass sich empirisch zwei hoch korrelierte, aber separierbare Konstrukte aufzeigen lassen. Da für beide Konstrukte dieselbe zentrale kognitive Operation angenommen wird (logisches Schlussfolgern; vgl. Abschnitt 3.1.1), sollte die Untersuchung des Zusammenhangs zudem als Evidenz für konvergente Validität dienen.

Das dritte Ziel bestand in der Untersuchung des Zusammenhangs zwischen PU und der Persönlichkeit der Testperson. Auf Grund der Konzeptualisierung von PU als Fähigkeit mit starken Bezügen zu einem klassischen Intelligenzkonstrukt wurden im Allgemeinen Zusammenhänge zwischen PU und der Persönlichkeit der Testperson erwartet, wie sie auch im Intelligenzbereich zu finden sind (z.B. Ackerman & Heggestad, 1997). Für PU und die in der Studie betrachteten Big Five-Faktoren bedeutet dies, dass im Sinne der diskriminanten Validität Zusammenhänge nahe Null erwartet wurden (vgl. Abschnitt 3.1.3). Darüber hinaus sollte die Rolle der tatsächlichen Ähnlichkeit zwischen Testperson und Zielperson bei der

Erfassung von PU untersucht werden. Wie in Abschnitt 3.1.3.1 dargestellt, gibt es Hinweise darauf, dass angenommene Ähnlichkeit zwischen Beurteiler:in und Zielperson mit akkurateren Persönlichkeitsbeurteilungen im Zusammenhang steht, sofern gleichzeitig tatsächliche Ähnlichkeit vorliegt (Funder et al., 1995; Paunonen & Hong, 2013; Vogt & Colvin, 2003). Auch wenn die empirische Evidenz gemischt ausfällt und diese Beziehung möglicherweise nur für die normative Komponente der Akkuratheit vorliegt (Human & Biesanz, 2011, 2012; Letzring, 2015; vgl. Abschnitt 3.1.3.1), ist ein Effekt auf die Erfassung von PU mit dem AcquA-Testdesign nicht auszuschließen. Zwar ist bei den AcquA-Aufgaben für die Aufgabenlösung nur individuelles Wissen über die Zielperson aus der Aneignungsphase erforderlich (vgl. Abschnitt 3.2), dennoch konnten Hellwig et al. (2020) zeigen, dass diejenigen AcquA-EU Items leichter waren, deren Emotionskontingenzen mit allgemeinem Wissen über die Regelhaftigkeiten von Emotionen übereinstimmten (Hellwig, 2016). Allgemeines Wissen scheint bei der Lösung der AcquA-Aufgaben somit eine gewisse Rolle zu spielen, zumindest wenn dieses mit den Iteminhalten übereinstimmt. Die Nutzung von allgemeinem Wissen und die Nutzung von Wissen über die eigene Persönlichkeit wurden wiederum als ähnliche Bestandteile des Persönlichkeitsbeurteilungsprozesses beschrieben (Paunonen & Hong, 2013; Vogt & Colvin, 2003). Ein gewisser Einfluss der Ähnlichkeit auf die Leistung bei den AcquA-PU Aufgaben kann somit nicht ausgeschlossen werden. Daher wurde die Annahme überprüft, dass ein höherer Zusammenhang zwischen PU und Persönlichkeit der Testperson vorliegt, wenn Zielperson und Testperson eine ähnliche Ausprägung auf der betrachteten Persönlichkeitseigenschaft aufweisen.

4.2 Methode

4.2.1 Konstruktion von AcquA-Aufgaben zur Erfassung von PU

Die AcquA-Aufgaben zur Erfassung von PU wurden nach den allgemeinen Prinzipen und Vorgaben des AcquA-Testdesigns von Schulze und Roberts (2015) konstruiert. Grundlegender Aufbau und Gestaltung der Aufgaben wurde zudem aus dem EAP-Test von Hellwig et al. (2020; vgl. auch Hellwig, 2016) übernommen und bei Bedarf an den neuen Inhaltsbereich angepasst. Die genaue Umsetzung und das gewählte allgemeine Aufgabendesign der AcquA-PU Aufgaben wird im Folgenden auf Basis der drei genannten Quellen beschrieben.

4.2.1.1 Ereignis-Verhaltens-Kontingenzen. Wie in Abschnitt 3.2.2 erläutert wurde, eignen sich für die Konstruktion von AcquA-Aufgaben zur Erfassung von PU Ereignis-

Verhaltens-Kontingenzen, mit denen die Persönlichkeitseigenschaften der Zielpersonen über deren typische Verhaltensweisen modelliert werden. Das typische Verhalten kann wiederum in Anlehnung an die Gestaltung von Hellwig et al. (2020) aus direkt beobachtbarem Verhalten sowie der verbalen Reaktion der Zielperson, die persönlichkeitsrelevante Äußerungen beinhalten kann, bestehen. Die Präsentation der Kontingenzen erfolgt ebenfalls wie bei Hellwig et al. in Form von Gesprächen und Interaktionen zwischen mehreren Personen. Folglich dienen die Kontingenzen in den AcquA-PU Aufgaben der Darstellung, mit welchem beobachtbaren Verhalten und welcher verbalen Reaktion die Zielperson auf bestimmte Ereignisse reagiert, die im Laufe der Interaktion auftreten. Darüber hinaus ist es möglich, dass Kontingenzen präsentiert werden, indem eine der interagierenden Personen direkt über das typische Verhalten einer Zielperson spricht (vgl. Hellwig et al., 2020).

In Abschnitt 3.2.2 wurde zudem erläutert, warum die Big Five-Faktoren geeignet sind, um die Persönlichkeitseigenschaften, die durch die Kontingenzen modelliert werden sollen, zu strukturieren und auszuwählen. Als Grundlage für die Erstellung der Kontingenzen wurden daher Items zur Erfassung der Big Five-Faktoren oder deren Facetten verwendet. So konnte, zumindest zu einem gewissen Ausmaß, sichergestellt werden, dass die Kontingenzen persönlichkeitsrelevantes Verhalten beinhalten. Hierfür ausgewählt wurden Items von Goldberg (1999) aus dem International Personality Item Pool (IPIP) zur Erfassung der Facetten des Abridged Big Five Dimensional Circumplex (AB5C)-Modells von Hofstee et al. (1992). Beim AB5C-Modell handelt es sich um ein Zirkumplex-Modell, das Persönlichkeitsfacetten durch die gleichzeitige Betrachtung von jeweils zwei der Big Five-Faktoren ableitet (Hofstee et al., 1992). Neben den Primärladungen von Items auf einem der Big Five werden im Modell explizit Höhe sowie Vorzeichen möglicher Sekundärladungen auf einem weiteren der Big Five berücksichtigt. Auf diese Weise ergeben sich 40 mögliche Misch-Facetten (Hofstee et al., 1992). Beispielsweise umfasst die Facette mit der Bezeichnung III+/I- vs. III-/I+ diejenigen Items, die eine positive Primärladung auf dem Faktor Gewissenhaftigkeit aufweisen sowie eine substantielle negative Sekundärladung auf der Extraversion.[5] Zu den 40 Misch-Facetten kommen zudem fünf weitere „pure-factor facets“ (Hofstee et al., 1992, S. 147; z.B. III+/III+ vs. III-/III-) hinzu, denen Items ohne substantielle Sekundärladung zugeordnet werden. Insgesamt resultieren somit 45 Facetten, neun je Big Five-Faktor (Hofstee et al., 1992). Das AB5C-Modell kann damit als informativer und umfassender angesehen werden als Modelle,

[5] Bei Hofstee et al. (1992) sind die Big Five-Faktoren folgendermaßen gekennzeichnet: Extraversion (I), Verträglichkeit (II), Gewissenhaftigkeit (III), Emotionale Stabilität (IV; positiver Pol von Neurotizismus), Offenheit für Erfahrungen (V; Intellekt bei Hofstee et al., 1992).

die eine Einfachstruktur anstreben (Goldberg & Velicer, 2006) und Items zur Erfassung der Facetten des AB5C-Modells ermöglichen eine der umfassendsten und differenziertesten Beschreibungen der Persönlichkeit (DeYoung et al., 2007). In der Realität kann nicht angenommen werden, dass das Verhalten und Erleben einer Person stets nur von einem der Big Five-Faktoren beeinflusst wird. Diese Annahme wird empirisch insofern gestützt, dass sich in Faktormodellen einschlägiger Inventare zur Erfassung der Big Five nach durchgeführter orthogonaler Rotation substantielle Nebenladungen von Facetten auf anderen Faktoren finden lassen (z.B. NEO-PI-R; Ostendorf & Angleitner, 2004). Dies ist ein Ergebnis, das sich auch auf Itemebene gezeigt hat (Goldberg & Velicer, 2006).

AB5C-Items bilden die Realität somit umfassender und differenzierter ab und stellen damit eine geeignete Grundlage für die Kontingenzen der AcquA-PU Aufgaben dar. In der Literatur werden allerdings auch Nachteile des AB5C-Modells beschrieben. So ist die Annahme des Modells, dass jeder der Big Five-Faktoren exakt neun Facetten aufweist, anzuzweifeln (Saucier & Ostendorf, 1999). Zudem resultieren durch die Kombination von jeweils zwei Faktoren auch inkonsistente Facetten, die inhaltlich wenig Sinn ergeben, sowie Facetten mit inhaltlich heterogenen Items (Saucier & Ostendorf, 1999). Und auch wenn die AB5C-Facetten deduktiv und unabhängig von spezifischen Items konzeptualisiert wurden (Hofstee et al., 1992; Woods & Anderson, 2016), erfolgt die empirische Zuordnung der Items zu den Facetten in der Regel mit Hilfe von Marker-Variablen der Big-Five (Hofstee et al., 1992), was zwangsläufig zu einer Abhängigkeit der Ergebnisse von der Auswahl der Marker-Variablen führt. All diese Nachteile wirken sich allerdings nicht auf die Nutzung der AB5C-Items für die Konstruktion der Kontingenzen aus, da das Scoring der AcquA-PU Aufgaben nicht von der Korrektheit des AB5C-Modells abhängt (vgl. Abschnitt 3.2).

Je Zielperson wurde mindestens ein IPIP-AB5C-Item (Goldberg, 1999) mit möglichst behavioralem Inhalt ausgewählt. Dieses stellte die Grundlage der für die Aufgabenlösung relevanten Kontingenz und damit das Scoring der Testpersonenantworten dar. Die restlichen IPIP-AB5C-Items wurden zudem zur Modellierung weiterer Reaktionen und Verhaltensweisen herangezogen, die auch als Kontingenzen in die Aufgabe eingebaut wurden, aber für die Aufgabenlösung irrelevant waren (vgl. Hellwig et al., 2020). Auf diese Art und Weise erfolgte die Modellierung der gesamten Interaktion zwischen den Personen einer AcquA PU-Aufgabe.

4.2.1.2 Format der Aufgabendarstellung. In den AcquA-EU Aufgaben von Hellwig et al. (2020) werden die Personen als grafische Avatare dargestellt und deren Interaktion mit Hilfe von Sprechblasen präsentiert. Darüber hinaus werden hier die emotionalen Reaktionen

ebenfalls als Textlabels dargestellt, sodass den Testpersonen sämtliche Information schriftlich dargeboten wird. Eine Übernahme dieser Aufgabendarstellung für die AcquA-PU Aufgaben hätte bedeutet, dass neben den Gesprächen zwischen den Personen auch das eigentlich beobachtbare Verhalten hätte schriftlich präsentiert werden müssen. Um Schlussfolgerungen über die Persönlichkeitseigenschaften anderer Personen möglichst alltagsnah abbilden zu können, wurde ein neues Format für die Darstellung der AcquA-PU Aufgaben verwendet, das eine direkte Beobachtung der Verhaltensweisen ermöglicht. Gewählt wurde anstelle des 2D-Format der AcquA-EU Aufgaben eine Darstellung mit Hilfe von 3D-Simulationen, die den Testpersonen als Videos präsentiert werden können. Dieses Format wurde gewählt, da 3D-Simulationen weiterhin die vollständige Kontrolle über Verhalten, Gestik und Mimik der Avatare ermöglichen.

4.2.1.3 Aneignungsphase. Analog zu den AcquA-EU Aufgaben von Hellwig et al. (2020) startet jede AcquA-PU Aufgabe mit einer kurzen verbalen Beschreibung der Situation und der dort involvierten Personen. Letztere werden zusätzlich mit Hilfe Portrait-ähnlicher Bilder vorgestellt, damit sie von den Testpersonen in den darauffolgenden Videosequenzen leichter identifiziert werden können. Anschließend wird ein Standbild des ersten Videoframes präsentiert, um den Testpersonen eine Orientierung in der Situation zu ermöglichen. Die Aneignungsphase kann schließlich selbstständig gestartet werden. Sofern diese aus mehreren Videosequenzen besteht, wird jede neue Sequenz wie oben dargestellt eingeführt. Jedes Video kann dabei nur einmal betrachtet werden.

In der Aneignungsphase einer AcquA-PU Aufgabe findet in der Regel eine Interaktion und Konversation zwischen mehreren Personen statt (vgl. Hellwig et al., 2020). Den Testpersonen wird vorab nicht mitgeteilt, welche Person eine Zielperson darstellt und am Ende der Aufgabe eingeschätzt werden muss und ebenso wird nicht mitgeteilt, welches Verhalten für die spätere Aufgabenlösung relevant ist (vgl. Hellwig et al., 2020; Schulze & Roberts, 2015). Die Aufgabe der Testpersonen besteht darin, die Interaktion zu beobachten und zu lernen, was die Personen als Reaktion auf bestimmte Ereignisse typischerweise tun und sagen. Die vorab definierten relevanten Kontingenzen einer oder mehrerer Zielpersonen werden innerhalb dieser Phase präsentiert und so sichergestellt, dass jede Testperson die Möglichkeit hat, Wissen über das für die spätere Aufgabenlösung relevante Verhalten der Zielpersonen zu erlangen. Wichtig ist hierbei, dass die Verhaltensweisen klar erkennbar und nicht auf Grund von Subtilität oder Schnelligkeit leicht zu übersehen sind, um die Verfügbarkeit der relevanten Information soweit wie möglich sicherzustellen (vgl. RAM; Funder, 1995).

Wie bei Hellwig (2016) sind die Avatare der AcquA-PU Aufgaben anhand des Aussehens (z.B. Kleidung, Haarfarbe) eindeutig unterscheidbar. Darüber hinaus bleibt das Aussehen der Avatare während der gesamten Aufgabe konstant, sodass weitestgehend sichergestellt werden kann, dass die Fähigkeit zur Wiedererkennung von Gesichtern (vgl. Soziales Gedächtnis als Teilfähigkeit der SI; Weis et al., 2006) keinen Einfluss auf die Leistung bei den Aufgaben hat. Auf Grund des in Abschnitt 3.2.2 beschriebenen Ziels, Emotionen soweit wie möglich auszuschließen, wurde bisher keine Animation der Mimik der Avatare vorgenommen. Emotionen, die über Gestik, Körperhaltung und Verhalten vermittelt werden, können auf Grund des Aufgabenformats allerdings nicht vollständig ausgeschlossen werden.

Durch ausschließlich einmalige Präsentation der Aneignungsphase wird wie bei Hellwig et al. (2020) sichergestellt, dass die Testpersonen während der Aufgabenlösung nicht mehr auf die dort präsentierte Information zugreifen können.

4.2.1.4 Anwendungsphase. Die Anwendungsphase startet mit einer kurzen verbalen Beschreibung der neuen Situation und der dort involvierten Personen (vgl. Hellwig et al., 2020). Je Anwendungsphase wird meist nur eine Zielperson fokussiert, sodass eine Aufgabe mehrere Anwendungsphasen umfassen kann. Es handelt sich dabei stets um eine Situation, die eine hinreichende Ähnlichkeit zur Situation der Aneignungsphase aufweisen soll (Schulze & Roberts, 2015). Diese Ähnlichkeit bezieht sich in erster Linie darauf, dass die neue Situation mit einem Ereignis endet, das aus der relevanten Kontingenz der Zielperson ableitbar ist. Die Ähnlichkeit muss somit insbesondere zwischen dem in der relevanten Kontingenz beschriebenen Ereignis aus der Aneignungsphase und dem präsentierten Ereignis in der Anwendungsphase sichergestellt werden.

Die Aufgabe der Testpersonen besteht zunächst erneut darin, die involvierten Personen zu beobachten. Nach Präsentation des relevanten Ereignisses folgt der zweite Teil der Anwendungsphase, in dem die Testpersonenantworten gesammelt werden. Den Testpersonen werden nacheinander mehrere mögliche Reaktionen der Zielperson – bestehend aus beobachtbarem Verhalten und verbaler Reaktion – auf das relevante Ereignis präsentiert. Diese möglichen Reaktionen müssen von den Testpersonen einzeln dahingehend eingeschätzt werden, wie wahrscheinlich oder unwahrscheinlich sie für die Zielperson wären. Hierfür steht eine 10-stufige Skala von *unwahrscheinlich* (0) bis *wahrscheinlich* (10) zur Verfügung, wobei nur die Skalenendpunkte beschriftet sind und die mittlere Position nicht ausgewählt werden kann (vgl. Hellwig et al., 2020). Die einzelnen möglichen Reaktionen bestehen aus kurzen Videoclips, die vollständig betrachtet werden müssen, bevor die Testperson eine endgültige

Einschätzung der Wahrscheinlichkeit vornehmen kann. Jede vorgenommene Einschätzung entspricht dabei einem Item der Aufgabe. Vor der Einschätzung werden die Testpersonen zudem explizit darauf hingewiesen, dass die Einschätzung auf dem in der Aneignungsphase erworbenen Wissen über die Zielperson basieren soll und dass sie davon ausgehen können, dass sie zuvor das typische Verhalten der Zielperson kennengelernt haben.

4.2.1.5 Scoring. In Studie 1 wurden wie bei Hellwig et al. (2020) ausschließlich deterministische Kontingenzen verwendet. In den entsprechenden Aufgaben wird somit angenommen, dass die Zielperson auf das in der Kontingenz beschriebene Ereignis *immer* mit der in der Kontingenz beschriebenen Verhaltensweise reagiert (Wahrscheinlichkeit von 1) und mit keiner anderen (Wahrscheinlichkeit von 0; Schulze & Roberts, 2015). Jede Kontingenz und jedes in einer Kontingenz definierte Ereignis wird in der Aneignungsphase einmalig präsentiert, sodass die Testpersonen nur diese und keine mit der Kontingenz inkonsistente Information erhalten (Hellwig et al., 2020). Folglich können die Testpersonen in der Anwendungsphase auch nur diese Information zur Einschätzung der präsentierten Reaktionen nutzen. Die einzuschätzenden Reaktionen entsprechen entweder der vorab definierten Kontingenz und können auf Basis der Information aus der Aneignungsphase durch logisches Schlussfolgern eindeutig als wahrscheinlich eingeschätzt werden (*likely-Items*; korrekte Antwort bei deterministischer Kontingenz: 10) oder sie sind nicht mit der Kontingenz zu vereinbaren und daher eindeutig unwahrscheinlich (*unlikely-Items*; korrekte Antwort bei deterministischer Kontingenz: 0; Schulze & Roberts, 2015). Die unlikely-Items können wiederum aus verschiedenen Gründen nicht mit der Kontingenz zu vereinbaren sein und entsprechend heterogen aussehen: In unlikely-Items wird entweder eine andere Ausprägung der fokussierten Persönlichkeitseigenschaft modelliert (z.B. niedrige anstatt hohe Extraversion) oder eine andere Persönlichkeitseigenschaft. Bei den likely-Items ist zudem wichtig, dass diese nicht exakt dasselbe beobachtbare Verhalten und nicht exakt dieselbe verbale Reaktion aus der Aneignungsphase beinhalten, da durch die Aufgaben primär logisches Schlussfolgern und nicht Merkfähigkeit angesprochen werden soll. Die Testpersonen werden im Rahmen der Instruktion zudem darauf hingewiesen, dass eine Reaktion der Zielperson nur dann als wahrscheinlich angesehen werden kann, wenn sowohl beobachtbares Verhalten als auch verbale Reaktion mit dem Verhalten der Zielperson aus der Aneignungsphase übereinstimmen (vgl. Hellwig, 2016).

Die Auswertung der AcquA-Aufgaben kann unter Verwendung von zwei Scoring-Methoden erfolgen: 1.) Dichotomes Scoring, 2.) Bestimmung von Distanzwerten (Schulze &

Roberts, 2015). Wie bei Hellwig et al. (2020) beschrieben, erhalten die Testpersonen beim dichotomen Scoring 1 Punkt, wenn das jeweilige Item korrekt als wahrscheinlich oder korrekt als unwahrscheinlich eingeschätzt wurde, das heißt wenn ein Skalenwert auf der richtigen Seite der Skala gewählt wurde. Wählen die Testpersonen einen Skalenwert auf der falschen Seite, erhalten sie für das Item 0 Punkte. Die Skala wird dabei in der Mitte (Skalenwert 5; nicht wählbar) geteilt und die Skalenwerte 0 bis 4 als unwahrscheinlich und die Skalenwerte 6 bis 10 als wahrscheinlich angesehen (Hellwig et al., 2020). Bei der Bestimmung von Distanzwerten entspricht der Itemscore der Abweichung zwischen gewähltem Skalenwert der Testperson und korrektem Skalenwert, der bei deterministischen Kontingenzen stets einen der Skalenendpunkte darstellt (Schulze & Roberts, 2015).

Beim Distanzscoring handelt es sich um einen Ansatz, der bisher insbesondere bei der Bewertung von Testpersonenantworten in Situational Judgment Tests (SJTs) angewendet wurde (z.B. Legree et al., 2005). Hierbei werden verschiedene Vorgehensweisen zur Bestimmung der Distanzwerte unterschieden: Einfache Distanz, quadrierte Distanz, einfache standardisierte Distanz und quadrierte standardisierte Distanz (Legree, 1995; Legree et al., 2005; Legree et al., 2010; MacCann et al., 2004; McDaniel et al., 2011). Bei den einfachen und quadrierten Distanzwerten entspricht der Itemscore der absoluten beziehungsweise quadrierten absoluten Abweichung zwischen gewähltem Skalenwert der Testperson und korrektem Skalenwert (MacCann et al., 2004). Bei den beiden standardisierten Distanzwerten werden gewählter und korrekter Skalenwert zunächst jeweils intraindividuell z-standardisiert, sodass über alle Items hinweg sowohl für die gewählten als auch für die korrekten Skalenwerte je Testperson $M = 0$ und $SD = 1$ gilt (MacCann et al., 2004; McDaniel et al., 2011). Ziel der z-Standardisierung ist die Kontrolle von Antworttendenzen (Legree, 1995).

Unter Verwendung von AcquA-EU Aufgaben hat ein Vergleich aller fünf Scoring-Methoden (d.h. Distanzwerte sowie das dichotome Scoring) ergeben, dass sich die unstandardisierten Distanzwerte nicht zur Auswertung der AcquA-Aufgaben eignen, da eine nicht interpretierbare zweifaktorielle Struktur der Aufgaben resultierte (Pisters & Schulze, 2017). Die insgesamt betrachtet besten Ergebnisse zeigten sich unter Verwendung der einfachen standardisierten Distanz (Pisters & Schulze, 2017), sodass diese Methode auch zur Auswertung der AcquA-PU Aufgaben verwendet werden sollte.

Im Zusammenhang mit den standardisierten Distanzwerten müssen allerdings auch potentielle Schwierigkeiten sowie die sehr begrenzte empirische Evidenz im geplanten Einsatzbereich berücksichtigt werden. So ist die absolute Höhe der Itemscores und damit auch des Testscores auf Grund der intraindividuellen z-Standardisierung nur eingeschränkt

interpretierbar. Darüber hinaus wurden standardisierte Distanzwerte bisher erst zweimal unter Verwendung eindeutig korrekter und zudem ausschließlich extremer Antworten (Skalenendpunkte) eingesetzt (Pisters & Schulze, 2017; Schulze & Jobmann, 2016). Der Einsatz beschränkt sich sonst eher auf Bereiche, in denen die korrekten Skalenwerte nicht eindeutig bestimmt werden können und daher mit Hilfe des (Experten) Consensus Scorings ermittelt werden (Legree et al., 2005; Weekley et al., 2006). Im Gegensatz zu den korrekten Skalenwerten der AcquA-Aufgaben mit deterministischen Kontingenzen variieren die korrekten Skalenwerte bei Verwendung des (Experten) Consensus Scorings über den gesamten Wertebereich der Skala. Zudem stellen sie keine konstanten Werte dar, sondern weisen eine eigene Verteilung auf (Legree et al., 2010). Zur Überprüfung der Stabilität der Ergebnisse unter Verwendung des standardisierten Distanzscorings wurde daher eine zusätzliche Auswertung der AcquA-PU Aufgaben mit Hilfe des dichotomen Scorings vorgenommen. Darüber hinaus wurden die AcquA-EU Aufgaben bisher überwiegend mit Hilfe des dichotomen Scorings ausgewertet (Hellwig et al., 2020), sodass nur durch zusätzliche Verwendung des dichotomen Scorings ein angemessener Vergleich zwischen den Ergebnissen der PU-Aufgaben und den Ergebnissen der EU-Aufgaben möglich war.

4.2.1.6 Umsetzung im 3D-Format. Die Umsetzung der AcquA-PU Aufgaben als 3D-Simulationen erfolgte mit Hilfe der Computersoftware iClone 6 (Reallusion, 2014) und hierfür zusätzlich erwerbbarem Content (https://marketplace.reallusion.com/iclone). Das Programm ermöglicht eine vollständige Kontrolle sowohl über die Gestaltung der Avatare und Szenarien als auch über Bewegungen und Mimik der Avatare. Praktisch mussten auf Grund der Komplexität des Programmes und des Aufwandes bei der Programmierung der Animationen allerdings Einschränkungen hingenommen werden.

Während die beobachtbaren Verhaltensweisen auch als solche in der 3D-Umgebung programmiert wurden, wurde zunächst auf eine Vertonung der Dialoge verzichtet. Grund hierfür war einerseits der in Abschnitt 3.2.2 genannte Ausschluss von Emotionen, um eine Abgrenzung zwischen PU und EU zu ermöglichen. Da Emotionen auch durch die Sprache vermittelt und erkannt werden (z.B. Banse & Scherer, 1996; Laukka et al., 2016; Scherer, 2003), sollte der Verzicht auf eine Vertonung diesen möglichen Einfluss von Emotionen ausschließen. Zum anderen wurde bisher noch keine auditive Präsentation von AcquA-Aufgaben vorgenommen. Die Veränderung des Aufgabenformates sollte sich daher zunächst auf einen Aspekt (3D-Simulation) konzentrieren und nicht zwei Aspekte auf einmal beinhalten. Die Dialoge wurden daher wie bei Hellwig et al. (2020) als Sprechblasen in die Aufgaben

integriert (vgl. Abbildung 4.1) und hierbei eine eindeutige Zuordnung zu den Avataren sichergestellt. Da die Testpersonen auf Grund der Video-basierten Präsentation die Geschwindigkeit der AcquA-PU Aufgaben nicht selbstständig kontrollieren können, musste vorab festgelegt werden, wie lange die Sprechblasen eingeblendet werden. Gewählt wurde eine großzügige Zeit, damit die Aufgabenleistung möglichst unabhängig von interindividuellen Unterschieden in der Lesegeschwindigkeit der Testpersonen erfasst wird. Nach Carver (1990, 1992) beträgt die Lesegeschwindigkeit durchschnittlich 433 ms pro Wort, wenn das Ziel das Merken und eine spätere Wiedergabe des Textinhaltes (inkl. spezifischer Fakten) ist. Da das Merken der verbalen Reaktion der Zielpersonen bei den PU-Aufgaben eine wichtige Rolle spielt, diente diese Angabe als Anhaltspunkt. Für eine bessere Umsetzbarkeit in iClone 6 wurde der Wert aufgerundet und somit für das Einblenden der Sprechblasen eine Zeit von 500 ms pro Wort verwendet. Brysbaerts (2019) kam auf Basis eines Reviews und einer Meta-Analyse zu dem Ergebnis, dass die meisten Erwachsenen bei nicht-fiktionalem Text eine Lesegeschwindigkeit von 175 bis 300 Wörtern pro Minute aufweisen, was etwa 200 bis 343 ms pro Wort entspricht. Das stützt die Annahme, dass mit 500 ms pro Wort eine großzügige Zeit gewählt wurde, die auch langsame Leser:innen nicht benachteiligt und zudem genügend Zeit zur Beobachtung der Verhaltensweisen lässt.

4.2.2 Stichprobe

Insgesamt wurden die Daten von 203 Personen erhoben. Eine Person musste auf Grund eines vorab definierten Sprachkriteriums (Personen mit einer anderen Muttersprache als Deutsch sollten seit mindestens 10 Jahren Deutsch sprechen) ausgeschlossen werden. Dieses Kriterium wurde auf Grund des hohen verbalen Anteils der eingesetzten Instrumente festgelegt. Für eine valide Erfassung der Konstrukte waren daher gute Deutschkenntnisse notwendig.

Die vollständige Analysestichprobe umfasste N = 202 Personen (153 Frauen, 49 Männer) im Alter von 17 bis 61 Jahren (M = 22.25, SD = 4.71; eine Person konnte auf Grund von technischen Problemen ihr Alter nicht angeben). Deutsch als Muttersprache gaben 179 Personen an und alle anderen Personen der Analysestichprobe sprachen seit mindestens 10 Jahren Deutsch. Die meisten Personen (n = 161) gaben als höchsten akademischen oder schulischen Abschluss die Fachhochschul- oder Hochschulreife (Abitur) an. Einen akademischen Abschluss (Bachelor, Master oder Diplom) wiesen 31 Personen auf, neun Personen einen Realschul- oder gleichwertigen Abschluss und eine Person befand sich zum Testzeitpunkt noch in schulischer Ausbildung. Der Großteil der Analysestichprobe (n = 160) wurde über Aushänge an der Bergischen Universität Wuppertal und Social Media rekrutiert

und bestand somit hauptsächlich aus Studierenden unterschiedlicher Fachrichtungen. Insgesamt 148 dieser Personen gaben an, noch zu studieren. Den größten Anteil an den Studierenden (n = 90) machten hierbei die Psychologiestudierenden aus. Darüber hinaus bestand die Stichprobe aus 42 Schüler:innen zweier Klassen eines Wuppertaler Berufskollegs.

4.2.3 Messinstrumente

Im Rahmen der ersten Studie sollte das AcquA-Testdesign sowohl in einem neuen Inhaltsbereich als auch unter Verwendung eines neuen Formats zur Aufgabendarstellung angewendet werden. Bei der Untersuchung des Zusammenhangs zwischen AcquA-Aufgaben zur Erfassung von PU und EU (vgl. Studienziel 2) musste daher folgende potentielle Problematik bedacht werden: Sollte sich eine empirische Separierbarkeit der Konstrukte zeigen, so ließe sich nicht eindeutig sagen, ob diese durch die Variation des Konstruktes (EU vs. PU), die Variation des Formats (2D vs. 3D) oder die Interaktion von Konstrukt und Format erzielt wurde. Daher war ein an eine Multitrait-Multimethod-Matrix (Campbell & Fiske, 1959) angelehntes Studiendesign erforderlich, in dem beide Konstrukte (EU und PU) durch AcquA-Aufgaben beider Formate (2D und 3D) erfasst werden.

4.2.3.1 Personality Understanding. Zur Erfassung von PU wurden insgesamt sechs neu konstruierte AcquA-PU Aufgaben eingesetzt, von denen auf Grund des angestrebten MTMM-Designs drei im alten 2D-Format und drei im neuen 3D-Format konstruiert wurden. Den relevanten Kontingenzen aller Aufgaben wurden IPIP-AB5C-Items von Goldberg (1999) zur Erfassung der Gewissenhaftigkeit zu Grunde gelegt. Hier spielen Emotionen in der Konzeptualisierung (McCrae & John, 1992) und folglich auch in der Operationalisierung kaum eine Rolle, sodass der Faktor geeignet erschien. Die Beschränkung auf Items nur eines Faktors erfolgte vor allem, um eine zu große inhaltliche Heterogenität in der ersten Studie zu vermeiden. Ausgewählt wurden Items der folgenden AB5C-Facetten der Gewissenhaftigkeit (Goldberg, 1999): Efficiency (AB5C-Kennzeichnung: III+/I+ vs. III-/I-), Orderliness (III+/V- vs. III-/V+) sowie Conscientiousness (III+/III+ vs. III-/III-).

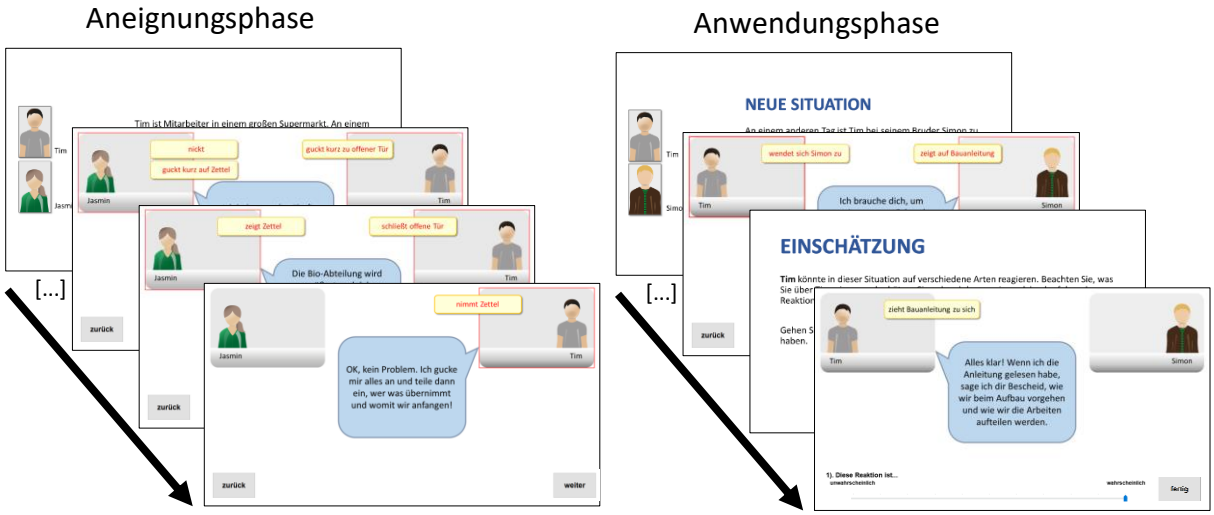
Die ausgewählten Items einer Facette dienten als Grundlage für jeweils zwei Aufgaben: Eine im 2D-Format und eine im 3D-Format. Die relevanten Kontingenzen dieser Aufgaben basieren also auf denselben Persönlichkeitsitems und sind damit vergleichbar. Das konkrete beobachtbare Verhalten und die verbale Reaktion der Zielpersonen sowie die Situation konnten sich zwischen den Aufgaben allerdings unterscheiden. Durch die Konstanthaltung der zugrundeliegenden Items sollte sichergestellt werden, dass gegebenenfalls auftretende

fehlende Zusammenhänge zwischen den Aufgaben im 2D- und 3D-Format auf den Wechsel im Aufgabenformat (vs. zu starke inhaltliche Heterogenität) zurückgeführt werden können.

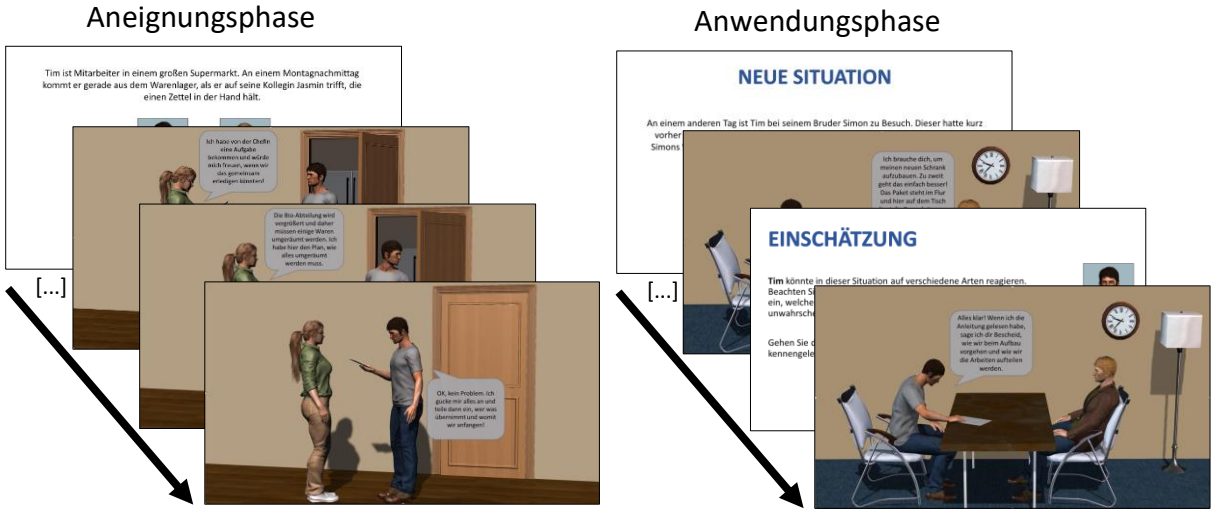
Abbildung 4.1

Beispielhafte Darstellung der AcquA-Aufgaben zur Erfassung von Personality Understanding im 2D-Format (Ausschnitte einer Aufgabe; Panel A) und 3D-Format (Screenshots aus den Videos; Panel B)

A:



B:



Die Konstruktion der Aufgaben im 2D-Format orientierte sich an einer grafisch überarbeiteten Version der AcquA-EU Aufgaben von Hellwig et al. (2020), die am Lehrstuhl für Methodenlehre und Psychologische Diagnostik der Bergischen Universität Wuppertal entwickelt wurde. Der einzige Unterschied zu den AcquA-EU Aufgaben besteht bei diesem

Format der AcquA-PU Aufgaben darin, dass anstelle der Emotionen das beobachtbare Verhalten schriftlich präsentiert wird, wobei neu auftretendes Verhalten in roter Schrift (inkl. roter Umrandung des Kästchens) erscheint (Hellwig et al., 2020). Beispiele der AcquA-PU Aufgaben im 2D- und 3D-Format in Form von Ausschnitten beziehungsweise Screenshots aus den Aufgaben sind in Abbildung 4.1 dargestellt. Eine Übersicht über verschiedene Aspekte der sechs neu konstruierten AcquA-PU Aufgaben ist zudem in Tabelle 4.1 zu finden.

Tabelle 4.1

Übersicht über verschiedene Aspekte der in Studie 1 eingesetzten AcquA-Aufgaben zur Erfassung von Personality Understanding (PU)

		Anzahl		
Aufgabe (Format)	AB5C-Facette der Kontingenzen	Zielpersonen in AP1	relevante Kontingenzen in AP1	Einschätzungen in AP2
PU1 (3D)	Efficiency	2	2	12
PU1 (2D)	Efficiency	2	2	12
PU2 (3D)	Orderliness	2	2	12
PU2 (2D)	Orderliness	2	2	12
PU3 (3D)	Conscientiousness	2	4	24
PU3 (2D)	Conscientiousness	2	3	18
Gesamt		12	15	90

Anmerkungen. AP1 = Aneignungsphase; AP2 = Anwendungsphase.

Analog zu der bereits existierenden Instruktion zu den AcquA-EU Aufgaben von Hellwig et al. (2020) wurde eine umfangreiche Instruktion für die AcquA-PU Aufgaben konstruiert. Die Testpersonen werden hier anhand einer kurzen, kommentierten Beispielaufgabe mit dem Aufgabenformat und der Aufgabenbearbeitung vertraut gemacht sowie genau instruiert, wie die Einschätzungen in den Anwendungsphasen vorzunehmen sind. Der vollständige Instruktionstext wurde den Testpersonen dabei zusätzlich auditiv präsentiert. Die Instruktion der AcquA-PU Aufgaben erfolgte zudem primär auf Basis des 3D-Formats. Das 2D-Format wurde in einem zweiten Teil der Instruktion als alternative Darstellung der Aufgaben eingeführt und die Beispielaufgabe zur Illustration im 2D-Format erneut präsentiert. Zusätzlich wurden die Teile der Instruktion, die die PU- von den EU-Aufgaben abgrenzen, grafisch hervorgehoben, um die Aufmerksamkeit der Testpersonen stärker auf die

Hauptcharakteristik der PU-Aufgaben zu lenken. Darüber hinaus wurden die Testpersonen darüber informiert, dass die Videos nur einmal betrachtet werden können.

4.2.3.2 Emotional Understanding. Für die Erfassung von EU wurden sechs bereits vorhandene AcquA-EU Aufgaben mit den Kennungen EU7, EU8, EU9, EU12, EU13 und EU15 eingesetzt. Diese Aufgaben wurden am Lehrstuhl für Methodenlehre und Psychologische Diagnostik der Bergischen Universität Wuppertal konstruiert und haben in vergangenen Studien gute psychometrische Eigenschaften gezeigt (Pisters & Schulze, 2017; Schulze & Jobmann, 2016). Auf Basis der Daten der Studie von Pisters und Schulze (2017) konnte zudem vorab die Eindimensionalität der Aufgaben überprüft und bestätigt werden.

Drei der Aufgaben wurden in ihrem ursprünglichen 2D-Format von den Testpersonen bearbeitet. Zur Umsetzung des angestrebten MTMM-Designs wurden die anderen drei Aufgaben mit Hilfe von iClone 6 (Reallusion, 2014) in ein 3D-Format überführt, wobei dies nach Möglichkeit ohne inhaltliche Veränderungen der Aufgaben vorgenommen wurde. Auf Grund der besseren Umsetzbarkeit wurden die Aufgaben EU8, EU9 und EU12 in das 3D-Format überführt und die Aufgaben EU7, EU13 und EU15 in ihrem ursprünglichen Format verwendet. Eine Übersicht über verschiedene Aspekte der sechs eingesetzten EU-Aufgaben ist in Tabelle 4.2 zu finden.

Tabelle 4.2

Übersicht über verschiedene Aspekte der in Studie 1 eingesetzten AcquA-Aufgaben zur Erfassung von Emotional Understanding (EU)

		Anzahl		
Aufgabe	Format	Zielpersonen in AP1	relevante Kontingenzen in AP1	Einschätzungen in AP2
EU7	2D	2	2	9
EU8	3D	2	2	8
EU9	3D	3	3	12
EU12	3D	3	3	18
EU13	2D	3	3	18
EU15	2D	3	3	16
Gesamt		16	16	81

Anmerkungen. AP1 = Aneignungsphase; AP2 = Anwendungsphase.

Bei den Aufgaben im 3D-Format erfolgte die Präsentation der Sprechblasen analog zu den AcquA-PU Aufgaben. Zudem wurden die Emotionslabels wie bei den AcquA-EU Aufgaben im 2D-Format ebenfalls schriftlich und beim erstmaligen Auftreten in roter Schrift (inkl. roter Umrandung des Kästchens) präsentiert. Die Zuordnung der Emotionslabels zu den Zielpersonen war dabei stets eindeutig. Die Präsentationsdauer der Sprechblasen wurde ebenfalls wie bei den AcquA-PU Aufgaben bestimmt, wobei je neu auftretender Emotion 1000 ms und je zuvor bereits präsentierter, anhaltender Emotion 500 ms zur Präsentationsdauer der Sprechblasen hinzuaddiert wurden.

Die AcquA-EU Aufgaben wurden mit Hilfe der von Hellwig et al. (2020) erstellten und am Lehrstuhl für Methodenlehre und Psychologische Diagnostik grafisch überarbeiteten Instruktion eingeleitet. Die Instruktion erfolgte primär auf Basis eines 3D-Formats dieser Instruktion. Das bereits vorhandene 2D-Format wurde wie bei den PU-Aufgaben in einem zweiten Teil der Instruktion als alternative Darstellungsvariante der Aufgaben eingeführt und das ursprüngliche Beispiel zur Illustration erneut im 2D-Design präsentiert. Zur Auswertung der AcquA-EU Aufgaben wurde ebenfalls primär das standardisierte Distanzscoring verwendet sowie zusätzlich das dichotome Scoring.

4.2.3.3 Persönlichkeit der Testpersonen. Im Rahmen des dritten Studienziels wurde die Annahme formuliert, dass ein höherer Zusammenhang zwischen PU und Persönlichkeit der Testperson vorliegt, wenn Zielperson und Testperson eine ähnliche Ausprägung auf der betrachteten Persönlichkeitseigenschaft aufweisen. Um eine Überprüfung dieser Annahme zu ermöglichen, musste die Persönlichkeit der Testpersonen möglichst vergleichbar zu der modellierten Persönlichkeit der Zielpersonen, sprich zu den modellierten relevanten Kontingenzen, erfasst werden. Daher wurde die Persönlichkeit der Testpersonen zum einen mit Hilfe der für die Aufgabenkonstruktion ausgewählten AB5C-Facetten der Gewissenhaftigkeit von Goldberg (1999) erfasst: Efficiency (III+/I+ vs. III-/I-; 11 Items, im Original $\alpha = .83$), Orderliness (III+/V- vs. III-/V+; 10 Items, $\alpha = .78$) sowie Conscientiousness (III+/III+ vs. III-/III-; 13 Items, $\alpha = .75$). Zur Einordnung und Kontrastierung der Zusammenhänge zwischen PU und den Facetten der Gewissenhaftigkeit wurden darüber hinaus zusätzlich drei AB5C-Facetten der Extraversion erhoben (Goldberg, 1999): Gregariousness (I+/I+ vs. I-/I-; 10 Items, $\alpha = .83$), Friendliness (I+/II+ vs. I-/II-; 10 Items, $\alpha = .85$) und Leadership (I+/V+ vs. I-/V-; 10 Items, $\alpha = .82$).

Zum Zeitpunkt der Studienplanung war keine deutsche Übersetzung der IPIP-AB5C-Items von Goldberg (1999) verfügbar, sodass eigene Übersetzungen angefertigt werden

mussten. Hierfür wurden zunächst unabhängige Übersetzungsvorschläge von zwei Psychologiestudent:innen sowie für einzelne Items bereits vorhandene Übersetzungen (Hartig et al., 2003; Ostendorf, n.d.; Schreiber & Iller, 2016) integriert. Im Anschluss wurden die Resultate von einer weiteren Person mit abgeschlossener Promotion im Fach Psychologie sowie sehr guten Englischkenntnissen geprüft und gegebenenfalls durch alternative Übersetzungsvorschläge ergänzt. Konnte bei Items keine unmittelbare Einigung auf eine Übersetzung erzielt werden, wurden die alternativen Übersetzungsvorschläge von vier weiteren Psycholog:innen mit mindestens Bachelorabschluss unabhängig voneinander in eine Rangfolge gebracht. Anschließend wurde die Übersetzung mit dem insgesamt besten Rang ausgewählt.

Die Verwendung eigener Übersetzungen erforderte zudem eine Überprüfung, ob sich die AB5C-Facettenstruktur von Goldberg (1999) auf Basis der deutschsprachigen Items replizieren lässt. Wie Goldberg et al. (2006) zu entnehmen ist, wurde die Facettenbildung bei Goldberg (1999) mittels des auf der IPIP-Website beschriebenen Vorgehens durchgeführt (Goldberg, n.d.). Hiernach erfolgte die Zuordnung der Items zu den AB5C-Facetten anhand der Korrelationen der einzelnen Items mit den Big Five-Faktoren, die wiederum mit Hilfe der 100 Marker-Adjektive von Goldberg (1992) erfasst wurden (vgl. Vorgehen von Hofstee et al., 1992). Die Überprüfung der Facettenstruktur der deutschsprachigen Items sollte auf eine ähnliche Weise unter Verwendung von deutschsprachigen Marker-Adjektiven erfolgen. Hierfür wurden Adjektive aus den umfangreichen lexikalischen Arbeiten von Ostendorf (1990) verwendet. Je Big Five-Faktor wurden diejenigen 20 Items mit den höchsten Faktorladungen ausgewählt (vgl. Ostendorf, 1990, Tabelle 52 bis 56), sodass sich ebenfalls 100 Marker-Adjektive ergaben. Diese wurden bereits von Saucier und Ostendorf (1999) als deutschsprachiges Pendant zu den 100 Marker-Adjektiven von Goldberg (1992) verwendet. Die Adjektive von Ostendorf (1990) wurden zudem verwendet, um die Ausprägung der Testpersonen auf allen Big Five-Faktoren zu erfassen.

Sowohl die IPIP-Items als auch die Marker-Adjektive sollten von den Testpersonen auf einer 6-stufigen Likert-Skala von *Ablehnung stark* (1) bis *Zustimmung stark* (6) beantwortet werden. Die übrigen Skalenpunkte waren ebenfalls beschriftet und alle Skalenpunkte wurden mit dem Ziel der Äquidistanz und einem einheitlichen Verständnis vorab näher erläutert.

4.2.4 Durchführung

Die vollständig computergestützte Datenerhebung wurde an der Bergischen Universität Wuppertal in einem ruhigen Laborraum durchgeführt. An einem Termin konnten bis zu sechs

Testpersonen gleichzeitig teilnehmen. Am Wuppertaler Berufskolleg wurden zwei Erhebungen mit jeweils einer gesamten Klasse in einem großen Computerraum durchgeführt. Anwesend waren hier mindestens zwei Versuchsleiter:innen sowie die jeweiligen Fachlehrer:innen, deren Unterrichtszeit in Anspruch genommen werden durfte.

Die Programmierung und Präsentation der Erhebungsinstrumente wurde mit Inquisit 4 (Millisecond Software, 2016) vorgenommen. Da an beiden Erhebungsorten keine einheitliche Monitorgröße realisiert werden konnte (Universität: 23.8 Zoll, Auflösung: 1920 x 1080 Pixel; Berufskolleg: 19 Zoll, Auflösung: 1280 x 1024 Pixel), wurde die Darstellung der Instrumente an die jeweilige Monitorgröße angepasst. So konnten Darstellungsunterschiede in der Größe der Videos und Grafiken weitestgehend ausgeglichen werden.

Nach umfangreicher und schriftlicher allgemeiner Information zur Studie gaben die Testpersonen ihre schriftliche Einwilligung in die Studienteilnahme (bei Minderjährigen wurde zusätzlich das Einverständnis einer erziehungsberechtigten Person eingeholt). Die Studie begann anschließend durch das Vorlesen der allgemeinen Instruktion, die den Testpersonen zusätzlich schriftlich am Computer präsentiert wurde. Schließlich konnte die Testung durch die Testpersonen am Computer selbstständig fortgesetzt werden. Die auditive Präsentation der Instruktionen zu den AcquA-PU und AcquA-EU Aufgaben erfolgte über einen bereitgelegten Kopfhörer, sodass individuelle Bearbeitungsgeschwindigkeiten realisiert werden konnten.

Die Reihenfolge der eingesetzten Erhebungsinstrumente unterschied sich je nach Erhebungsort. Der Ablauf der Untersuchung an der Universität kann Tabelle 4.3 entnommen werden. Bei den AcquA-PU und AcquA-EU Aufgaben konnten Reihenfolgeeffekte nicht ausgeschlossen werden. Aus diesem Grund wurden zwei Reihenfolgen programmiert, zu denen die Testpersonen randomisiert zugewiesen wurden: Bei Reihenfolge A wurden im ersten Block zuerst die PU-Aufgaben und nach der Pause im zweiten Block die EU-Aufgaben bearbeitet. Bei Reihenfolge B erfolgte dies genau umgekehrt. Auch innerhalb eines AcquA-Blocks konnten Reihenfolgeeffekte der einzelnen Aufgaben nicht ausgeschlossen werden. Innerhalb des AcquA-PU und AcquA-EU Blocks wurde die Reihenfolge der Aufgaben daher unvollständig ausbalanciert. Hierbei wurde darauf geachtet, dass Aufgaben im 2D- und 3D-Format abwechselnd präsentiert werden. Zudem sollte stets mit einer Aufgabe im 2D-Format begonnen werden, um den Testpersonen den Einstieg in die Aufgabenbearbeitung zu erleichtern. Bei den PU-Aufgaben wurde außerdem darauf geachtet, dass diejenigen Aufgaben, die auf denselben Persönlichkeitsitems basierten, mit insgesamt maximalem Abstand zueinander präsentiert werden. Die Adjektive sowie die IPIP-Items wurden jeweils in randomisierter Reihenfolge bearbeitet.

Tabelle 4.3

Reihenfolge und Zeitschätzung (in Minuten) der in Studie 1 eingesetzten Instrumente (Erhebungsort: Bergische Universität Wuppertal)

Instrumente Reihenfolge A	Instrumente Reihenfolge B	Zeitschätzung
1. Allgemeine Instruktion & Demografie	1. Allgemeine Instruktion & Demografie	7
2. AcquA-PU	2. AcquA-EU	65
3. Big Five Marker-Adjektive	3. Big Five Marker-Adjektive	10
Pause (ca. 5 Minuten; optional)		
4. AcquA-EU	4. AcquA-PU	65
5. IPIP-AB5C-Items	5. IPIP-AB5C-Items	10

Anmerkung. Insgesamt betrug die geschätzte Studiendauer 162 Minuten (inkl. Pause).

Für die Erhebung am Berufskolleg wurde die Reihenfolge der Instrumente leicht abgewandelt. Im Vorfeld wurde vermutet, dass auf Grund der begrenzt verfügbaren Zeit nicht alle Testpersonen die Studie bis zum Ende durchführen können. Aus diesem Grund wurden diejenigen Instrumente an das Studienende gesetzt, bei denen fehlende Werte die potentiell geringsten Auswirkungen auf das Erreichen der Studienziele hatten. Der Hauptfokus der Studie lag auf den AcquA-Aufgaben, wohingegen der Zusammenhang zwischen PU und der Persönlichkeit der Testperson ein eher untergeordnetes Ziel darstellte. Daher wurden die Marker-Adjektive sowie die IPIP-AB5C-Items an das Studienende gesetzt. Die Pause wurde zudem nach dem ersten AcquA-Block angeboten.

Nach Abschluss der Erhebung erhielten die Testpersonen in beiden Fällen eine Aufwandsentschädigung in Höhe von 12 €. Studierende der Bergischen Universität Wuppertal konnten alternativ auch eine Gutschrift von Versuchspersonenstunden entsprechend der tatsächlichen Studiendauer erhalten. Insgesamt variierte diese zwischen 1:50 h und 3:00 h.

4.2.5 Statistische Analysen

4.2.5.1 Allgemeines. Für die Überprüfung der faktoriellen Validität der AcquA-PU Aufgaben, der faktoriellen Struktur der AcquA-EU Aufgaben sowie für die Untersuchung der Zusammenhänge zwischen PU und EU sowie PU und den Big Five-Faktoren wurden konfirmatorische Faktorenanalysen (CFAs) verwendet. In den jeweiligen Modellen wurden

sowohl für PU als auch für EU nicht die Itemscores der AcquA-Aufgaben, sondern Aufgabenparcels als Indikatoren verwendet. Diese wurden gebildet, indem nach abgeschlossener Itemselektion alle zu einer Aufgabe gehörigen Items durch Mittelwertbildung zusammengefasst wurden. Wie Hellwig et al. (2020) ausführlich erläutern, ist auf Grund des Aufbaus der AcquA-Aufgaben (mehrere Einschätzungen beziehen sich auf dieselbe Aneignungsphase) nicht davon auszugehen, dass die Items lokale stochastische Unabhängigkeit aufweisen. Entsprechende Analysen haben dies für die EU-Aufgaben auch bereits bestätigt (Hellwig et al., 2020). Bei den PU-Aufgaben muss daher ebenfalls davon ausgegangen werden, dass bei Verwendung der einzelnen Items im Messmodell Kovarianzen zwischen den Fehlern vorliegen (Lee et al., 2001). Eine Berücksichtigung dieser Fehlerkovarianzen im Modell oder die Verwendung eines Bifaktor-Modells mit PU als Generalfaktor und den Aufgaben als spezifische Faktoren war auf Grund der für die vorliegende Anzahl an Items (vgl. Tabelle 4.1 und 4.2) relativ kleinen Stichprobe allerdings nicht möglich. Daher wurden wie bei Hellwig et al. (2020) in allen Modellen Aufgabenparcels als Indikatoren verwendet.

Für die Schätzung der CFA-Modelle wurde sowohl Mplus (Muthén & Muthén, 2017) als auch das R-Paket Latent Variable Analysis (lavaan; Rosseel, 2012) verwendet. Reliabilitäten wurden mit Hilfe des R-Pakets Methods for the Behavioral, Educational, and Social Sciences (MBESS; Kelley, 2007, 2020) geschätzt. Bei vorliegender Eindimensionalität und falls nicht anders angegeben, wurden Cronbachs α sowie McDonalds ω bestimmt sowie als Konfidenzintervalle für beide Koeffizienten Bootstrap-Perzentil-Intervalle mit 10000 Replikationen (vgl. Empfehlungen von Kelley, 2020; Kelley & Pornprasertmanit, 2016).

4.2.5.2 Itemselektion AcquA-Aufgaben. Auf Grund der Neukonstruktionen der AcquA-PU Aufgaben war eine Selektion geeigneter Items erforderlich. Diese wurde getrennt für die beiden verwendeten Scoring-Methoden in einem iterativen Prozess mit Hilfe aufgabenweiser CFAs durchgeführt. Die Ziele der Selektion waren die Herstellung von Eindimensionalität der einzelnen Aufgaben sowie die Elimination psychometrisch ungeeigneter Items. Geschätzt wurde daher jeweils ein eindimensionales Modell, in dem die einzelnen Items der jeweiligen Aufgabe als Indikatoren verwendet wurden. Eindimensionalität wurde angestrebt, da die selektierten Items für die anschließenden Analysen zu Aufgabenparcels zusammengefasst werden sollten (vgl. oben) und fehlende Eindimensionalität der Parcels Probleme mit sich bringen kann. Beispielsweise können multidimensionale Parcels

verzerrte Messmodelle und Schätzungen der Ladungen zur Folge haben, was wiederum die Interpretationen der latenten Variablen sowie Zusammenhänge erschwert (Little et al., 2002).

Die Itemselektionen wurden unter Berücksichtigung von Vorzeichen und Höhe der standardisierten Faktorladungen der Items sowie der durch die Elimination erzielten Verbesserung im Modell-Fit vorgenommen. Folgende Selektionskriterien führten dabei in der angegebenen Rangfolge zur schrittweisen Elimination von Items:

1. Vorliegen einer negativen Faktorladung (beginnend bei der größten negativen Ladung), da hierdurch ein eindeutig ungeeignetes Leistungstestitem anzeigt wird
2. Vorliegen einer standardisierten Faktorladung < .10 (beginnend bei der niedrigsten Ladung), was einer Varianzaufklärung durch den Faktor von < 1 % entspricht
3. Größte Verbesserung im Modell-Fit bis zum Erreichen vorab festgelegter Cut-Off-Kriterien

Nach jeder Elimination eines Items wurden die Selektionskriterien erneut in der angegebenen Rangfolge geprüft und anschließend gegebenenfalls das nächste Item eliminiert. Zur Evaluation des Modell-Fits wurden Punktschätzung und 90%-Konfidenzintervall des Root Mean Square Error of Approximation (RMSEA) betrachtet sowie der Comparative Fit Index (CFI). Konnten durch die Elimination von verschiedenen Items gleich gute oder sehr ähnliche RMSEA- und CFI-Werte erzielt werden, wurden zudem die χ^2-Teststatistik sowie die Höhe der standardisierten Ladungen mit in die Selektionsentscheidung einbezogen. Die Cut-Off-Werte für RMSEA und CFI wurden wie folgt festgelegt: Der RMSEA sollte einen Wert $\leq .05$ annehmen (Browne & Cudeck, 1992) und zusätzlich sollte das 90%-Konfidenzintervall den Wert 0 beinhalten. Beim CFI wurde der zu erreichende Cut-Off auf $\geq .95$ festgelegt (Hu & Bentler, 1999). Beendet wurde die Itemselektion, sobald alle standardisierten Ladungen Werte $\geq .10$ aufwiesen sowie die Cut-Off-Kriterien für den Modell-Fit erreicht wurden.

Da die Itemscores bei Verwendung des standardisierten Distanzscorings von dem für die z-Standardisierung verwendeten Itempool abhängen, wurden die Itemscores nach jeder Elimination sowie vor jeder Schätzung des Modell-Fits auf Basis eines reduzierten Itempools (d.h. ohne alle bereits eliminierten Items sowie ohne das potentiell zu eliminierende Item) neu bestimmt. Nach Abschluss der aufgabenweisen Itemselektionen wurden die standardisierten Distanzwerte auf Basis des finalen Itempools (d.h. ohne die eliminierten Items aller Aufgaben) neu bestimmt, die Messmodelle der einzelnen Aufgaben anhand der Selektionskriterien erneut überprüft und gegebenenfalls weitere Selektionsschritte zur erneuten Erreichung der Kriterien durchgeführt. Bei Verwendung des dichotomen Scorings war dies nicht erforderlich.

Die Itemselektion wurde unter Verwendung eines eigens auf die AcquA-Aufgaben angepassten und um das standardisierte Distanzscoring erweiterten R-Pakets durchgeführt.[6] Dieses Paket basiert auf dem R-Paket Subtests Using Algorithmic Rummaging Techniques (stuart) von Schultze (2018) und greift für die Modellschätzung auf Mplus (Muthén & Muthén, 2017) zurück.

Die beschriebene Itemselektion wurde nicht nur bei den AcquA-PU Aufgaben, sondern auch bei den AcquA-EU Aufgaben vorgenommen. Die Gründe hierfür waren, dass bei den verwendeten Aufgaben bisher noch keine finale Itemselektion durchgeführt und zudem durch die Übertragung von drei Aufgaben in das 3D-Format relevante Änderungen vorgenommen wurden.

4.2.5.3 Zusammenhang PU und EU. Im Anschluss an die Itemselektionen wurde bei PU und EU die angenommene Eindimensionalität der Aufgaben überprüft sowie der Zusammenhang auf latenter Ebene mit Hilfe eines zweifaktoriellen CFA-Modells geschätzt (vgl. Studienziel 2). Zur Überprüfung, ob es sich bei PU und EU um empirisch separierbare Konstrukte handelt, wurde das zweifaktorielle mit einem einfaktoriellen Modell verglichen.

Darüber hinaus wurde das zweifaktorielle Modell zur Untersuchung des Einflusses des Aufgabenformats (2D vs. 3D) zu einem MTMM-Modell erweitert und geschätzt. Die Aufnahme beider Aufgabenformate als zwei verschiedene, aber korrelierte Methodenfaktoren hätte die Verwendung eines Correlated Trait Correlated Method (CTCM)-Modells erfordert (Eid et al., 2006), mit dem Identifikations-, Schätz- und Interpretationsprobleme einhergehen (Eid et al., 2006, Eid et al., 2003). Erschwerend kam hinzu, dass in der vorliegenden Studie nur zwei Traits und zwei Methoden verwendet wurden. Zur Lösung einiger Probleme des CTCM-Modells wurde von Eid (2000) das CTC(M-1)-Modell vorschlagen, bei dem ein Methodenfaktor weniger modelliert wird, als Methoden verwendet wurden. Dieses liegt zudem als erweitertes Multiple-Indicator Model vor (Eid et al., 2003) und eignete sich daher für das vorliegende Untersuchungsdesign, in dem drei Indikatoren pro Trait-Methoden-Einheit vorlagen. Das CTC(M-1)-Modell basiert auf der Idee, eine Methode als Vergleichsmethode festzulegen und die andere(n) Methode(n) mit dieser Vergleichsmethode zu kontrastieren (Eid, 2000). Im Fall von zwei verwendeten Methoden wird somit nur ein Methodenfaktor modelliert. Dieser stellt einen Residualfaktor dar und erfasst die (messfehlerfreien) Abweichungen der

[6] Für die umfangreiche Unterstützung bei der technischen Umsetzung der geplanten Itemselektion sowie des standardisierten Distanzscorings in R möchte ich mich an dieser Stelle herzlich bei Markus Jansen bedanken.

Indikatorwerte von den Werten, die auf Basis der Vergleichsmethode zu erwarten gewesen wären (Eid, 2000; Eid et al., 2003).

Die Wahl der Vergleichsmethode soll nach Eid (2000) aus theoretischen Überlegungen und Forschungsinteressen heraus erfolgen. Da in der Vergangenheit ausschließlich das 2D-Format verwendet wurde, wurde dieses als Vergleichsmethode festgelegt. Im CTC(M-1)-Modell für multiple Indikatoren können für die Methoden zudem Trait-spezifische Methodenfaktoren modelliert und anhand der Höhe der Korrelation zwischen diesen Methodenfaktoren die Generalisierbarkeit des Methodeneffekts über die Traits hinweg geprüft werden (Eid et al., 2003). Auf Grund der hohen methodischen Überlappung zwischen den PU- und EU-Aufgaben wurden im vorliegenden Fall keine Trait-spezifische Methodenfaktoren modelliert beziehungsweise deren Korrelation auf 1 gesetzt. Eine grafische Darstellung des spezifizierten CTC(M-1)-Modell kann Abbildung A1 in Anhang A entnommen werden.

4.2.5.4 Zusammenhang PU und Persönlichkeit der Testperson. Um den Zusammenhang zwischen PU und der mit den IPIP-AB5C-Items erfassten Persönlichkeit der Testpersonen zu bestimmen, musste zunächst die Skalenzuordnung der übersetzten Items überprüft und gegebenenfalls eine neue Skalenbildung auf Basis des AB5C-Modells vorgenommen werden. Die Skalenbildung erfolgte in Anlehnung an die Vorgehensweise von Hofstee et al. (1992), wobei einige Modifikationen vorgenommen wurden.

Bei der ursprünglichen Vorgehensweise, die im Folgenden auf Basis von Hofstee et al. (1992; vgl. auch Woods & Anderson, 2016) beschrieben wird, werden zunächst eine Hauptkomponentenanalyse mit orthogonaler Varimax-Rotation der Big Five-Marker durchgeführt und die Faktorwerte der Big Five ermittelt. Anschließend werden die IPIP-AB5C-Items mit den Faktorwerten der Big Five korreliert und die Korrelationen als Faktorladungen der Items auf den Big Five interpretiert. Auf die so erhaltene Faktorladungsmatrix wird schließlich der AB5C-Algorithmus angewendet, bei dem die Items entsprechend ihrer Primär- und Sekundärladungen einer der 45 AB5C-Facetten zugeordnet werden. Dabei wird ein Item dann einer der fünf „pure-factor facets“ (Hofstee et al., 1992, S. 147) zugeordnet, wenn der absolute Wert der Primärladung mindestens 3.73-fach höher ist als der absolute Wert der Sekundärladung. Alle anderen Items werden einer der 40 Misch-Facetten zugeordnet.

Folgende Modifikationen wurden vorgenommen: Bei den in der vorliegenden Arbeit verwendeten Big Five-Markern von Ostendorf (1990) handelt es sich um Adjektive, die im Jahr 1990 veröffentlicht wurden. Auf Grund von erwartbaren Veränderungen in der deutschen Sprache konnte nicht davon ausgegangen werden, dass sich die Adjektive in gleicher Weise

wie vor über 30 Jahren als Marker eignen, das heißt weiterhin die gleichen psychometrischen Eigenschaften aufweisen. Darüber hinaus befanden sich unter den 20 Adjektiven pro Faktor inhaltlich redundante Items und solche, die die beiden Poole einer Dimension darstellen (z.B. ungesellig und gesellig; Ostendorf, 1990). In einem ersten Schritt wurde daher eine Itemselektion vorgenommen. Hierzu wurde je Big Five-Faktor ein eindimensionales CFA-Modell geschätzt. Als Indikatoren wurden jeweils die 20 Adjektive verwendet, die bei Ostendorf (1990) dem Faktor zugeordnet wurden. Die Itemselektion wurde mit dem Ziel der Eindimensionalität der Adjektive eines Faktors sowie der Auswahl der geeignetsten Items vorgenommen und analog zu dem Vorgehen bei den AcquA-Aufgaben durchgeführt. Zusätzlich wurde der Inhalt der Items mit in die Selektionsentscheidungen einbezogen, um die Anzahl inhaltlich redundanter Items zu reduzieren. Darüber hinaus wurden die Faktorwerte der Big Five nicht wie bei Hofstee et al. (1992) unter Verwendung der Hauptkomponentenanalyse bestimmt, da es sich hierbei nicht um eine echte Faktorenanalyse handelt und die Hauptkomponenten keine latenten Variablen darstellen (Fabrigar et al., 1999). Alternativ wurden die Faktorwerte mit Hilfe eines fünffaktoriellen CFA-Modells in Mplus geschätzt. Zudem wurden die Faktoren im CFA-Modell nicht als orthogonal, sondern als korreliert angenommen, da Studien auf Basis verschiedener Instrumente übereinstimmend Interkorrelationen der Big Five aufgezeigt haben (z.B. John et al., 2008; Schulze & Roberts, 2006; Soto & John, 2017; van der Linden et al., 2010) und die Annahme der Orthogonalität somit nicht angemessen erschien. Die so gewonnenen Faktorwerte wurden anschließend mit den IPIP-AB5C-Items korreliert und der beschriebene AB5C-Algorithmus auf die Korrelationsmatrix angewandt.

Zur Überprüfung der im Rahmen des dritten Studienziels formulierten Annahme, dass ein höherer Zusammenhang zwischen PU und Persönlichkeit der Testperson vorliegt, wenn beide eine ähnliche Ausprägung auf der betrachteten Persönlichkeitseigenschaft aufweisen, wurden die selektierten Items der AcquA-PU Aufgaben hypothesengeleitet zu Parcels zusammengefasst. Es wurden diejenigen Items der 3D- und 2D-Aufgaben zusammengefasst, die sich auf eine Zielperson mit derselben relevanten Kontingenz beziehen, das heißt deren Scoring-Rationale dasselbe IPIP-AB5C-Item zu Grunde liegt. Die so gebildeten Parcels wurden schließlich mit den gebildeten IPIP-AB5C-Skalen korreliert. Auf diese Weise konnte die Höhe der Korrelationen geschätzt werden, wenn bei Testperson und Zielperson dieselbe Persönlichkeitseigenschaft betrachtet wird (d.h. wenn IPIP-AB5C-Skala und PU-Parcel dasselbe Item beinhalten bzw. sich auf dasselbe Item beziehen).

Darüber hinaus wurden die selektierten Big Five-Marker verwendet, um in einem sechsfaktoriellen Modell die latenten Zusammenhänge zwischen PU und allen Big Five-Faktoren zu schätzen.

4.2.5.5 Power. Für die zentralen Analysen der Studie wurde die Power auf Basis der zur Verfügung stehenden Stichprobengröße bestimmt. Zum einen wurde die Power im Zusammenhang mit dem Messmodell für PU (und somit auch für EU) betrachtet. Die Bestimmung erfolgte unter Annahme eines eindimensionalen Messmodells mit sechs Indikatoren für den χ^2-Test des absoluten Fits mit H_0: RMSEA = 0, α = .05, angenommener Effekt: RMSEA = .10, df = 9, einem N von 197 und wurde mit Hilfe des von Preacher und Coffman (2006) publizierten R-Skript Generators vorbereitet. Die Poweranalyse in R ergab ein Ergebnis von 1-β = .85. Darüber hinaus erfolgte mit Hilfe des Monte Carlo Ansatzes in Mplus (Muthén & Muthén, 2002, 2017) die Bestimmung der Power für den ebenfalls im Fokus der Studie stehenden Zusammenhang zwischen PU und EU auf Basis eines zweifaktoriellen CFA-Modells (10000 Replikationen). Für die Schätzung wurden standardisierte Ladungen von .66 bei EU (durchschnittliche Ladung der sechs verwendeten EU-Aufgaben [standardisiertes Distanzscoring] bei Pisters & Schulze, 2017) und von .55 bei PU (konservative Annahme auf Grund der Neukonstruktionen) angenommen. Auf Basis der im zweiten Studienziel formulierten Annahme zum Zusammenhang zwischen PU und EU wurde dieser auf .50 festgelegt, was in einer Power von 1 resultierte (N = 197). Beide Poweranalysen deuten somit darauf hin, dass eine ausreichend große Stichprobe untersucht wurde.

4.3 Ergebnisse

Bei drei Personen der Analysestichprobe traten technische Probleme auf, sodass die Aufgabe EU9 nicht bearbeitet werden konnte. Zudem musste die Erhebung bei drei weiteren Personen vorzeitig abgebrochen werden. Je nach Analyse variiert die Stichprobengröße daher zwischen N = 197 bis 201 Personen.

Im Folgenden werden die Ergebnisse präsentiert, die bei Verwendung des standardisierten Distanzscorings zur Auswertung aller AcquA-PU und AcquA-EU Aufgaben ermittelt wurden. Die detaillierten Ergebnisse auf Basis des dichotomen Scorings befinden sich in Anhang B. Näher eingegangen wird auf diese nur vereinzelt und insbesondere, wenn beide Scoring-Methoden zu abweichenden Ergebnissen geführt haben.

4.3.1 AcquA-PU Aufgaben: Itemanalysen und Itemselektion

Schiefe, Kurtosis sowie Histogramme der Itemscores der AcquA-PU Aufgaben deuteten darauf hin, dass bei keinem der Items eine univariate Normalverteilung vorlag (Range Schiefe: -1.76 bis 5.33; 46 der 90 Items wiesen eine Schiefe > 2 auf; Range Kurtosis: -1.39 bis 33.65) und folglich bei keiner Aufgabe von einer multivariaten Normalverteilung der Items ausgegangen werden konnte. Für die CFA-Modelle der Itemselektion wurde daher ein Maximum Likelihood (ML)-Schätzer mit gegenüber Verletzungen der Annahme einer multivariaten Normalverteilung robusten Standardfehlern sowie χ^2-Teststatistik verwendet (MLM-Schätzer mit Satorra-Bentler χ^2; Muthén & Muthén, 2017). Auch wenn die für die Itemselektion vorab festgelegten Cut-Off-Kriterien des Modell-Fits (vgl. Abschnitt 4.2.5.2) für den nicht robusten ML-Schätzer vorgeschlagen wurden (Browne & Cudeck, 1992; Hu & Bentler, 1999), wurden diese beibehalten. Zum einen standen keine gesonderten Kriterien für den MLM-Schätzer zur Verfügung und zum anderen sprechen die Ergebnisse von Yu (2002) dafür, dass die gewählten Kriterien auch bei Verwendung des MLM-Schätzers akzeptable Ergebnisse liefern.

Auf Basis der aufgabenweisen CFAs wurde die Itemanzahl von 90 auf 64 reduziert. Deskriptive Statistiken der selektierten Items sind Tabelle A1 in Anhang A zu entnehmen. Die part-whole korrigierten Trennschärfen der selektierten Items, berechnet über alle Items hinweg, reichen von .14 bis .65 (M = .43, SD = .12) und die Itemmittelwerte von 0.21 bis 1.16 (M = 0.42, SD = 0.19), wobei bei Letzteren auf Grund der standardisierten Distanzen höhere Werte eine höhere Schwierigkeit anzeigen. Tabelle A1 in Anhang A ist zudem zu entnehmen, dass bei Verwendung des dichotomen Scorings zum Teil andere Items eliminiert wurden als beim standardisierten Distanzscoring. Aus der Tabelle geht ebenfalls hervor, dass bei Aufgabe PU1o keine Überschneidungen der bei beiden Scoring-Methoden selektierten Itemmengen vorliegen.

Die selektierten Items wurden anschließend durch Mittelwertbildung zu aufgabenweisen Parcels sowie einem PU-Gesamtscore zusammengefasst. Die deskriptiven Statistiken der gebildeten Parcels und des Gesamtscores finden sich in Tabelle 4.4.

Die Ergebnisse der Parcel-Analysen bei Verwendung des dichotomen Scorings sind in Tabelle B1 in Anhang B dargestellt. Dieser Tabelle ist vor allem zu entnehmen, dass beim dichotomen Scoring insgesamt eine ähnliche Anzahl an Items eliminiert wurde sowie dass die PU-Aufgaben – abgesehen von Aufgabe PU1o – extrem leicht waren.

Tabelle 4.4

Deskriptive Statistiken der Personality Understanding (PU)-Parcels sowie des gesamten AcquA-PU Tests (PU_{ges}) nach der Itemselektion (standardisiertes Distanzscoring)

Parcel[a]	Itemanzahl[b]	*M*	*SD*	Schiefe[c]	Kurtosis[d]
PU1n	9 (12)	0.40	0.24	1.48	4.57
PU1o	6 (12)	0.50	0.34	1.56	3.40
PU2n	7 (12)	0.40	0.32	1.85	4.35
PU2o	10 (12)	0.48	0.26	0.65	0.67
PU3n	18 (24)	0.39	0.20	0.80	1.67
PU3o	14 (18)	0.40	0.22	0.51	0.05
PU_{ges}	64 (90)	0.42	0.20	0.75	1.17

Anmerkungen. N = 201. Die verwendete Statistik zur Schätzung der Kurtosis nimmt beim Vorliegen einer Normalverteilung den Wert 0 an.

[a] o = 2D-Format, n = 3D-Format. [b] Angabe in Klammern bezieht sich auf die Anzahl vor der Itemselektion. [c] SE = 0.17. [d] SE = 0.34.

4.3.2 AcquA-PU Aufgaben: Messmodell und Reliabilität

Auf Grund der nicht anzunehmenden univariaten und damit auch nicht anzunehmenden multivariaten Normalverteilung der Indikatoren (vgl. Tabelle 4.4) wurde für das einfaktorielle CFA-Modell der AcquA-PU Aufgaben erneut der robuste MLM-Schätzer verwendet. Das angenommene Modell zeigte mit χ^2 = 18.72, df = 9, p = .03, CFI = .98, RMSEA = .07, 90% CI = [.02, .12] einen akzeptablen bis guten Fit. Die standardisierten Faktorladungen, die für die Aufgaben im 2D- und 3D-Format ähnlich hoch ausfielen, sind in Tabelle 4.5 dargestellt.

Die Schätzung der Reliabilität erfolgte auf Grund der überwiegend bestätigten Eindimensionalität unter Verwendung von Cronbachs α sowie McDonalds ω auf Basis der sechs PU-Parcels. Cronbachs α für den Gesamtwert liegt bei .85, 95% CI = [.80, .88], McDonalds ω für den PU-Faktor ebenfalls bei .85, 95% CI = [.80, .88].[7]

[7] Die Bestimmung des Koeffizienten ω_H, der sich auch dann eignet, wenn das eindimensionale Modell nicht perfekt auf die Daten passt (Kelley & Pornprasertmanit, 2016), führte zu derselben Punktschätzung wie ω. Auch das 95%-Konfidenzintervall fiel mit den Grenzen .80 bis .89 nur leicht breiter aus.

Tabelle 4.5

Ergebnis der konfirmatorischen Faktorenanalyse der Personality Understanding (PU)-Parcels (standardisiertes Distanzscoring)

Parcel[a]	standardisierte Faktorladung
PU1n	.75***
PU1o	.53***
PU2n	.69***
PU2o	.69***
PU3n	.90***
PU3o	.82***

Anmerkungen. $N = 201$.

[a] o = 2D-Format, n = 3D-Format.

***$p < .001$.

Unter Verwendung des dichotomen Scorings resultierte ein ähnlicher Modell-Fit. Die standardisierten Ladungen sowie die Reliabilitätsschätzungen fielen allerdings durchweg und zum Teil deutlich geringer aus (Range standardisierte Ladungen: .23 bis .65; $\alpha = .62$, $\omega = .62$; vgl. Tabelle B2 und Abschnitt B.2 in Anhang B).

4.3.3 AcquA-EU Aufgaben: Itemanalysen und Itemselektion

Auf Basis von Schiefe, Kurtosis sowie Histogrammen der Itemscores des standardisierten Distanzscorings konnte bei den AcquA-EU Aufgaben ebenfalls keine multivariate Normalverteilung angenommen werden (Range Schiefe: -1.09 bis 3.89, vier Items wiesen eine Schiefe > 2 auf; Range Kurtosis: -1.31 bis 20.48). Für die CFA-Modelle der Itemselektion wurde daher der MLM-Schätzer verwendet.

Das in der Methode beschriebene Vorgehen der Itemselektion führte bei Aufgabe EU12 zur Elimination von 14 der ursprünglich 18 Items. Von den verbliebenen vier Items wiesen zwei im einfaktoriellen CFA-Modell der Aufgabe standardisierte Ladungen unter .20 auf, die anderen beiden Items hingegen sehr hohe standardisierte Ladungen von .63 und .92. Um zu überprüfen, ob eine zu starke Gewichtung der beiden letztgenannten Items im Selektionsprozess zu der großen Anzahl eliminierter Items geführt hat, wurde die Itemselektion leicht abgewandelt wiederholt. In dieser zweiten Selektion wurden zuerst die beiden Items mit den sehr hohen Ladungen eliminiert und anschließend nach dem ursprünglichen Vorgehen

weiter selektiert. Auf diese Weise wurden 12 der 18 Items eliminiert. Die verbliebenen sechs Items wiesen im CFA-Modell der Aufgabe standardisierte Ladungen zwischen .13 und .60 auf. Um eine Entscheidung für eine der beiden Itemselektionen zu treffen, wurden die im Anschluss an die Itemselektionen geschätzten Messmodelle für EU und die jeweilige Ladung für das Parcel der Aufgabe EU12 betrachtet. Die standardisierte Ladung des Parcels der ersten Selektion betrug .29, die des Parcels der zweiten Selektion .60. Der Fit beider Modelle war fast identisch. Darüber hinaus ergab sich auf Basis der Ergebnisse der zweiten Selektion eine höhere Reliabilität (α = .70 vs. .78; ω = .70 vs. 78). Insgesamt führte somit die zweite, leicht angepasste Itemselektion der Aufgabe EU12 zu besseren Ergebnissen, sodass diese für das weitere Vorgehen verwendet wurde.

Insgesamt wurde auf Basis der aufgabenweisen CFAs die Itemanzahl von ursprünglich 81 auf 39 reduziert. Tabelle A2 in Anhang A sind die deskriptiven Statistiken der Items nach der Itemselektion zu entnehmen. Die part-whole korrigierten Trennschärfen der selektierten Items, berechnet über alle Items hinweg, liegen bei .09 bis .56 (*M* = .33, *SD* = .12), die Itemmittelwerte bei 0.29 bis 1.33 (*M* = 0.65, *SD* = 0.24). Die deskriptiven Statistiken der nach der Itemselektion gebildeten Aufgaben-Parcels und eines EU-Gesamtscores finden sich in Tabelle 4.6.

Tabelle 4.6

Deskriptive Statistiken der Emotional Understanding (EU)-Parcels sowie des gesamten AcquA-EU Tests (EU_{ges}) nach der Itemselektion (standardisiertes Distanzscoring)

Parcel[a]	Itemanzahl[b]	*M*	*SD*	Schiefe[c]	Kurtosis[d]
EU7o	4 (9)	0.60	0.35	0.71	0.70
EU8n	4 (8)	0.53	0.37	1.41	1.75
EU9n	8 (12)	0.54	0.25	0.23	-0.39
EU12n	6 (18)	0.68	0.30	0.25	0.02
EU13o	10 (18)	0.72	0.27	0.20	0.06
EU15o	7 (16)	0.77	0.35	0.63	0.56
EU_{ges}	39 (81)	0.65	0.21	0.13	-0.16

Anmerkungen. N = 198. Die verwendete Statistik zur Schätzung der Kurtosis nimmt beim Vorliegen einer Normalverteilung den Wert 0 an.

[a] o = 2D-Format, n = 3D-Format. [b] Angabe in Klammern bezieht sich auf die Anzahl vor der Itemselektion. [c] *SE* = 0.17. [d] *SE* = 0.34.

Auch bei den EU-Items ist Tabelle A2 in Anhang A zu entnehmen, dass bei beiden Scoring-Methoden teilweise unterschiedliche Items selektiert wurden (siehe insbesondere Items der Aufgabe EU12). Die deskriptiven Statistiken der EU-Parcels unter Verwendung des dichotomen Scorings sind in Tabelle B3 in Anhang B dargestellt. Auffällig sind hier die beiden extrem leichten Aufgaben EU8 und EU9.

4.3.4 AcquA-EU Aufgaben: Messmodell und Reliabilität

Zur Überprüfung der faktoriellen Struktur der AcquA-EU Aufgaben wurde entsprechend den Annahmen ein einfaktorielles CFA-Modell unter Verwendung der Aufgaben-Parcels als Indikatoren geschätzt. Eine Betrachtung von Schiefe, Kurtosis sowie Histogrammen der Parcels deutete auf Abweichungen der Indikatoren von einer univariaten Normalverteilung hin (insbesondere Parcel EU8n; vgl. Tabelle 4.6). Die Ergebnisse des Shapiro-Wilk-Tests auf Normalverteilung (vgl. Tabelle A3 in Anhang A) bestätigte dies für mehrere Indikatoren, sodass erneut der robuste MLM-Schätzer verwendet wurde. Das angenommene Modell zeigte mit $\chi^2 = 5.13$, $df = 9$, $p = .82$, CFI = 1, RMSEA = 0, 90% CI = [.00, .05] einen perfekten Fit. Die standardisierten Faktorladungen sind in Tabelle 4.7 zu finden und fallen auch hier im Schnitt für die Aufgaben im 2D- und 3D-Format ähnlich hoch aus. Cronbachs α für den EU-Gesamtwert beträgt .78, 95% CI = [.73, .82], McDonalds ω für den EU-Faktor ebenfalls .78, 95% CI = [.73, .83].

Beim dichotomen Scoring resultierte ein deskriptiv schlechterer, aber immer noch guter Modell-Fit. Die standardisierten Ladungen der Parcels im CFA-Modell sowie die Reliabilitätsschätzungen fielen allerdings auch hier überwiegend und teilweise deutlich geringer aus (Range standardisierte Faktorladungen: .33 bis .70; $\alpha = .64$, $\omega = .65$; vgl. Tabelle B4 und Abschnitt B.4 in Anhang B).

Tabelle 4.7

Ergebnis der konfirmatorischen Faktorenanalyse der Emotional Understanding (EU)-Parcels (standardisiertes Distanzscoring)

Parcel[a]	standardisierte Faktorladung
EU7o	.66***
EU8n	.57***
EU9n	.72***
EU12n	.60***
EU13o	.60***
EU15o	.58***

Anmerkungen. N = 198.

[a] o = 2D-Format, n = 3D-Format.

***p < .001.

4.3.5 Zusammenhang PU und EU

Das zweifaktorielle CFA-Modell mit den Faktoren PU und EU unter Verwendung der Aufgaben-Parcels als Indikatoren zeigte eine sehr gute Passung auf die Daten (MLM-Schätzer): χ^2 = 63.77, df = 53, p = .15, CFI = .99, RMSEA = .03, 90% CI = [.00, .06]. Die standardisierten Faktorladungen sowie die Faktorkorrelation, die mit einem Wert von .66 signifikant von 1 verschieden ist (α = .05; vgl. Konfidenzintervall), sind in Tabelle 4.8 zu finden. Ein alternatives einfaktorielles Modell zeigte mit χ^2 = 160.95, df = 54, p < .001, CFI = .86, RMSEA = .10, 90% CI = [.08, .12] (standardisierte Faktorladungen siehe Tabelle 4.8) deskriptiv einen schlechteren Fit als das zweifaktorielle Modell (ΔCFI = -.13, ΔRMSEA = -.07). Ein für den MLM-Schätzer vorgeschlagener χ^2-Differenzentest unter Verwendung einer skalierten Teststatistik nach Satorra und Bentler (2010) wurde mit Hilfe des R-Pakets lavaan (Rosseel, 2012) durchgeführt. Dieser ergab mit $\bar{T}_d$ = 95.80, df = 1, p < .001 eine auf einem α-Niveau von .001 signifikant bessere Passung des zweifaktoriellen Modells gegenüber dem einfaktoriellen Modell.

Tabelle 4.8

Standardisierte Faktorladungen der Personality Understanding (PU)-Parcels und Emotional Understanding (EU)-Parcels für das zweifaktorielle sowie das einfaktorielle CFA-Modell (standardisiertes Distanzscoring)

	zweifaktorielles Modell		einfaktorielles Modell
Parcel[a]	PU	EU	PEU
PU1n	.76***		.76***
PU1o	.53***		.50***
PU2n	.68***		.66***
PU2o	.70***		.71***
PU3n	.89***		.87***
PU3o	.85***		.84***
EU7o		.63***	.43***
EU8n		.57***	.42***
EU9n		.76***	.62***
EU12n		.57***	.38***
EU13o		.60***	.49***
EU15o		.57***	.44***
Faktorkorrelation [95% CI]			
PU - EU	.66 [.53, .78]		-

Anmerkungen. N = 197. CFA = Konfirmatorische Faktorenanalyse. CI = Konfidenzintervall.

[a] o = 2D-Format, n = 3D-Format.

****p* < .001.

Im nächsten Schritt wurde das beschriebene CTC(M-1)-Modell geschätzt. Wie in Abbildung A1 (Anhang A) dargestellt, wurde für alle PU- und EU-Aufgaben im 3D-Format eine Ladung auf dem Methodenfaktor spezifiziert. Das Modell zeigte zunächst keine Konvergenz. Erst nach Entfernung von neun multivariaten Ausreißern auf Basis der in Mplus berechneten Mahalanobis-Distanz ($p \leq .001$) konvergierte das Modell und zeigte mit $\chi^2 = 71.40$, $df = 47$, $p = .01$, CFI = .97, RMSEA = .05, 90% CI = [.03, .08] einen guten Fit (MLM-Schätzer). Die standardisierten Ladungen auf dem Methodenfaktor liegen bei -.26 bis .18 und zeigen kein systematisches Ladungsmuster (vgl. Tabelle 4.9). Die standardisierten Ladungen auf den Faktoren für PU und EU sowie die Faktorkorrelation zwischen PU und EU

weisen im Vergleich zu den Ergebnissen des Modells ohne Methodenfaktor nur geringe Unterschiede auf (vgl. Tabelle 4.8 und 4.9). Ein χ^2-Differenzentest unter Verwendung der skalierten Teststatistik von Satorra und Bentler (2010) ergab mit $\bar{T}_d$= 7.35, *df* = 1, *p* = .29 keine signifikant bessere Passung des CTC(M-1)-Modells gegenüber einem zweifaktoriellen Modell ohne Methodenfaktor und mit demselben Stichprobenumfang (ΔCFI = -.00, ΔRMSEA = -.00).[8]

Tabelle 4.9

Standardisierte Faktorladungen der Personality Understanding (PU)-Parcels und Emotional Understanding (EU)-Parcels für das CTC(M-1)-Modell (standardisiertes Distanzscoring)

	Faktor		
Parcel[a]/Faktor	PU	EU	Methodenfaktor (MF)
PU1n	.80***		.18
PU1o	.44***		
PU2n	.68***		-.07
PU2o	.66***		
PU3n	.90***		-.19
PU3o	.86***		
EU7o		.60***	
EU8n		.50***	-.23
EU9n		.78***	-.26*
EU12n		.55***	-.04
EU13o		.57***	
EU15o		.57***	
Faktorkorrelation [95% CI]			
EU	.61 [.49, .73]		
MF	0[b]	0[b]	

Anmerkungen. N = 188. CTC(M-1) = Correlated Trait Correlated Method Minus One; CI = Konfidenzintervall.

[a] o = 2D-Format, n = 3D-Format. [b] Faktorkorrelation wurde auf 0 fixiert.

*p < .05. ***p < .001.

[8] Für den χ^2-Differenzentest wurde ein zweifaktorielles Modell unter Ausschluss der multivariaten Ausreißer verwendet. Modell-Fit (*N* = 188): χ^2 = 78.48, *df* = 53, *p* = .01, CFI = .97, RMSEA = .05, 90% CI = [.02, .07].

Unter Verwendung des dichotomen Scorings ergaben sich für den Zusammenhang zwischen PU und EU nur leicht abweichende Ergebnisse. So fiel der Zusammenhang zwischen beiden Faktoren mit .74 deskriptiv etwas höher aus als beim standardisierten Distanzscoring (vgl. Tabelle B5 in Anhang B). Das CTC(M-1)-Modell konvergierte erst nach dem Setzen von Restriktionen und zeigte durchweg positive standardisierte Ladungen der Parcels auf dem Methodenfaktor zwischen .14 und .25 (vgl. Tabelle B6 in Anhang B). Zu beachten ist bei den Ergebnissen des dichotomen Scorings allerdings, dass die Modell-Fits dieser Analysen nicht zufriedenstellend ausfielen.

4.3.6 PU und Persönlichkeit der Testperson

4.3.6.1 Skalenbildung IPIP-AB5C-Items. Wie in Abschnitt 4.2.5.4 beschrieben, wurde für die Bildung der AB5C-Skalen zunächst eine Selektion der Marker-Adjektive von Ostendorf (1990) vorgenommen. Die Betrachtung von Schiefe, Kurtosis sowie Histogrammen der Adjektive resultierte in der Verwendung des MLM-Schätzers für die eindimensionalen CFA-Modelle der Itemselektion (Range Schiefe: -2.01 bis 0.69; Range Kurtosis: -1.31 bis 4.34). Auf Basis der faktorweisen CFAs wurde die Adjektivanzahl von 100 auf 53 reduziert. Die finale Adjektivauswahl, Fit-Statistiken der finalen einfaktoriellen CFA-Modelle sowie Reliabilitätsschätzungen je Big Five-Faktor finden sich in Tabelle A4 in Anhang A.

Im Anschluss an die Itemselektionen der Adjektive wurden das fünffaktorielle CFA-Modell mit korrelierten Faktoren geschätzt (MLM-Schätzer, $\chi^2 = 125.80$, $df = 80$, $p < .001$, CFI = .96, RMSEA = .05, 90% CI = [.04, .07]) und die Faktorwerte der Personen auf den Big Five-Faktoren mit Mplus geschätzt. Als Indikatoren wurden auf Grund der für die Itemanzahl verhältnismäßig kleinen Stichprobe nicht die Adjektive an sich verwendet, sondern Adjektiv-Parcels. Je Big Five-Faktor wurden drei Parcels gebildet, indem die selektierten eindimensionalen Adjektive zufällig und gleichmäßig auf die Parcels aufgeteilt und durch Mittelwertbildung zusammengefasst wurden (Little et al., 2002). Als Güte der Schätzungen der Faktorwerte wurden die von Mplus ausgegebenen Werte für die Korrelation zwischen den geschätzten Faktorwerten und den Faktoren (Muthén & Muthén, 2017) betrachtet, die bei .90 (Neurotizismus) bis .95 (Extraversion) lagen. Laut Gorsuch (2008) sollten diese Korrelationen mindestens .80 betragen, sodass die geschätzten Faktorwerte als für die weiteren Analysen geeignet beurteilt werden können. Die Faktorwerte wurden schließlich mit den IPIP-AB5C-Items von Goldberg (1999) korreliert und im Anschluss wurde mit Hilfe des AB5C-Algorithmus von Hofstee et al. (1992) eine Skalenbildung vorgenommen. Um die Stabilität der

Skalenbildung und die Abhängigkeit des Ergebnisses von den gebildeten Parcels zu überprüfen, wurde die zufällige Aufteilung der Adjektive auf die Parcels zwei weitere Male wiederholt, die Faktorwerte jeweils neu geschätzt und die Skalenbildung der IPIP-AB5C-Items erneut vorgenommen. Der Fit des fünffaktoriellen CFA-Modells fiel bei den Parcel-Sets 2 und 3 deskriptiv etwas schlechter aus als bei Parcel-Set 1 (Parcel-Set 2: $\chi^2 = 203.62$, $df = 80$, $p < .001$, CFI = .90, RMSEA = .09, 90% CI = [.07, .10]; Parcel-Set 3: $\chi^2 = 165.51$, $df = 80$, $p < .001$, CFI = .93, RMSEA = .07, 90% CI = [.06, .09]). Für die Güte der Faktorwerte resultierten bei Parcel-Set 2 Werte zwischen .88 (Neurotizismus) und .95 (Gewissenhaftigkeit) und bei Parcel-Set 3 zwischen .89 (Offenheit bzw. Intellekt) und .95 (Extraversion).

Weder für die IPIP-AB5C-Items der Gewissenhaftigkeit noch für die der Extraversion ergaben sich eindeutige Zuordnungen zu den AB5C-Skalen. Dies äußerte sich in wenigen Fällen durch geringe Unterschiede zwischen der höchsten und zweithöchsten Korrelation der Items mit den Faktorwerten. Darüber hinaus zeigte sich bei einer großen Anzahl an Items keine eindeutig zweithöchste Korrelation und somit keine eindeutige Sekundärladung auf den Big Five. Hier gab es zwischen der zweit-, dritt- und zum Teil vierthöchsten Korrelation zu den Faktorwerten nur geringfügige Unterschiede. Über die drei Sets der Faktorwerte hinweg äußerte sich dies zudem in einer nicht stabilen Skalenzuordnung einiger Items. Für die finalen Zuordnungen der Items zu den AB5C-Skalen wurden daher die Ergebnisse aus allen drei Faktorwerte-Sets zusammen betrachtet und diejenige Skala gewählt, die bei mindestens zwei der drei Sets auf Basis des AB5C-Algorithmus resultierte.

Für den Faktor Gewissenhaftigkeit wurden auf diese Weise fünf AB5C-Skalen gebildet: 1.) III+/III+ vs. III-/III-, 2.) III+/I+ vs. III-/I-, 3.) III+/II+ vs. III-/II-, 4.) III+/IV+ vs. III-/IV- und 5.) III+/V+ vs. III-/V-. Die Skala III+/III+ vs. III-/III- bestand lediglich aus zwei Items und wurde daher nicht weiter betrachtet. Zudem konnte ein Item keiner Skala zugeordnet werden, da hier die höchste Korrelation nicht zur Gewissenhaftigkeit vorlag. Die Zuordnungen der einzelnen Items zu allen fünf AB5C-Skalen finden sich in Tabelle A5 in Anhang A und deskriptive Statistiken sowie Reliabilitätsschätzungen der vier im Folgenden verwendeten Skalen in Tabelle A6 in Anhang A. Die Reliabilitätsschätzungen liegen zwischen $\omega_H = .68$ und .87 und nur eine Skala zeigte einen Wert unter .70.

Für den Faktor Extraversion wurden drei AB5C-Skalen gebildet: 1.) I+/I+ vs. I-/I-, 2.) I+/II+ vs. I-/II- und 3.) I+/V+ vs. I-/ V-. Ein Item zeigte die höchste Korrelation zur Gewissenhaftigkeit und wurde entsprechend der zweithöchsten Korrelation der Skala III+/I+ vs. III-/I- zugewiesen. Die Zuordnung aller Extraversions-Items zu den AB5C-Skalen findet

sich in Tabelle A7 in Anhang A und deskriptive Statistiken sowie Reliabilitätsschätzungen in Tabelle A6 in Anhang A. Die Reliabilitätsschätzungen liegen zwischen ω_H = .67 und .91 und wie bei der Gewissenhaftigkeit weist nur eine Skala einen Wert unter .70 auf.

Sowohl bei den Gewissenhaftigkeits-Skalen als auch bei den Extraversions-Skalen müssen die Reliabilitätsschätzungen auf Basis von Cronbachs α wegen der nicht sichergestellten Eindimensionalität mit Vorsicht interpretiert werden. Aus diesem Grund wurde auch ω_H nach Kelley und Pornprasertmanit (2016) anstelle von McDonalds ω bestimmt. Dieser eignet sich auch dann, wenn das eindimensionale Modell nicht perfekt auf die Daten passt und weitere spezifische Faktoren vorliegen (Kelley & Pornprasertmanit, 2016).

4.3.6.2 Zusammenhang PU-Parcels und AB5C-Skalen. Zur Überprüfung der Annahme, dass ein höherer Zusammenhang zwischen PU und Persönlichkeit der Testperson vorliegt, wenn Ziel- und Testperson eine ähnliche Ausprägung auf der betrachteten Persönlichkeitseigenschaft aufweisen, wurden die selektierten Items der PU-Aufgaben hypothesengeleitet zu spezifischen Parcels zusammengefasst (vgl. Abschnitt 4.2.5.4). Schließlich wurden die so gebildeten Parcels mit den zuvor gebildeten AB5C-Skalen korreliert. Deskriptive Statistiken und Schätzungen der Reliabilität der PU-Parcels[9] befinden sich in Tabelle 4.10, die Ergebnisse der Korrelationsanalysen in Tabelle 4.11. In Tabelle 4.11 wurden diejenigen Korrelationen fett markiert, bei denen auf Grund von Ähnlichkeit zwischen Testperson und Zielperson ein höherer Zusammenhang erwartet werden konnte. Beispielsweise wurde dem Scoring der Items im Parcel P3_c24r ein IPIP-Item zu Grunde gelegt, das Teil der Skala III+/V+ vs. III-/V- ist. Parcel P3_c24r bezieht sich also auf die Leistung beim Schlussfolgern über Zielpersonen, deren Persönlichkeit (basierend auf dem IPIP-AB5C-Item) sich zu einem gewissen Ausmaß mit Skala III+/V+ vs. III-/V- deckt. Zu beachten ist bei der Interpretation der Korrelationen, dass bei den PU-Parcels auf Grund des standardisierten Distanzscorings höhere Werte eine schlechtere Leistung anzeigen.

[9] Zur Bildung der hypothesengeleiteten PU-Parcels wurden Items von zwei Aufgaben (2D-Format und 3D-Format) zusammengefasst (vgl. Abschnitt 4.2.5.4), bei denen aufgabenspezifische Abhängigkeiten und damit korrelierte Fehler zu erwarten sind. Aus diesem Grund wurde erneut ω_H anstelle von ω bestimmt (vgl. Kelley & Pornprasertmanit, 2016).

Tabelle 4.10

Deskriptive Statistiken und Reliabilitätsschätzungen (ω_H; Kelley & Pornprasertmanit, 2016) der hypothesengeleitet gebildeten Personality Understanding (PU)-Parcels (standardisiertes Distanzscoring)

Parcel	Itemanzahl	*M*	*SD*	Schiefe[a]	Kurtosis[b]	ω_H [95% CI][c]
P1_c11	9	0.39	0.25	1.56	4.38	.65 [.47, .76]
P1_c14r	6	0.52	0.28	0.77	0.69	.42 [.25, 57]
P2_c15r	9	0.42	0.31	1.44	2.44	.73 [.56, .83]
P2_c12	8	0.47	0.24	0.43	0.26	.53 [.37, .62]
P3_c1	10	0.40	0.22	0.36	0.33	.66 [.56, .72]
P3_c32	9	0.38	0.23	0.76	0.60	.73 [.63, .79]
P3_c26r	5	0.46	0.27	0.62	0.80	.45 [.31, .57]
P3_c24r	8	0.37	0.23	0.99	1.81	.68 [.56, .75]

Anmerkungen. N = 201. CI = Konfidenzintervall. Aufbau Parcelname: P1 bis P3 = Aufgabennummer, c = Gewissenhaftigkeit, zweite Zahl = Nummer des IPIP-Items der Kontingenz, r = Zielperson besitzt geringe Ausprägung auf dem Faktor Gewissenhaftigkeit.
[a] SE = 0.17. [b] SE = 0.34. [c] Bei Schätzung der Konfidenzintervalle für ω_H traten wiederholt Schätzprobleme auf (z.B. Nicht-Konvergenz des Modells, negative Varianzen).

Die Ergebnisse des dichotomen Scorings finden sich in Tabelle B7 in Anhang B. Insgesamt resultierten unter Verwendung beider Scoring-Methoden maximal kleine Korrelationen, die kein eindeutiges und nicht das erwartbare Muster aufweisen. Der einzige erwartungskonforme Zusammenhang, der sich bei beiden Scoring-Methoden ergab, ist der zwischen Parcel P3_c24r (IPIP-Item der Kontingenz: „Do not plan ahead“; Goldberg, 1999) und der AB5C-Skala III+/V+ vs. III-/V-. Die Korrelation von r = .15 (standardisiertes Distanzscoring) beziehungsweise r = -.16 (dichotomes Scoring) ist so zu interpretieren, dass je geringer die Ausprägung der Testperson auf dieser Skala ausfiel, desto besser konnte sie über die wenig gewissenhaften Zielpersonen des Parcels schlussfolgern. Auf Grund des ansonsten unsystematischen Korrelationsmusters und der Menge an geschätzten Korrelationen sollte diese eine Korrelation aber nicht überbewertet werden. Darüber hinaus fielen die Reliabilitätsschätzungen je nach PU-Parcel sehr unterschiedlich und zum Teil sehr niedrig aus (vgl. Tabelle 4.10 sowie Tabelle B7 in Anhang B).

Tabelle 4.11

Korrelationen der hypothesengeleitet gebildeten Personality Understanding (PU)-Parcels (standardisiertes Distanzscoring) mit den Skalen des Abridged Big Five Dimensional Circumplex (AB5C)-Modells

	AB5C-Skalen						
PU-Parcel	III+/I+	III+/II+	III+/IV+	III+/V+	I+/I+	I+/II+	I+/V+
P1_c11	.01	-.05	-.03	**.06**	-.18**	-.15*	-.07
P1_c14r	.08	-.02	**.04**	.13	-.15*	-.11*	-.05
P2_c15r	**.03**	-.18*	-.02	.01	-.07	-.02	.06
P2_c12	**.09**	-.08	.08	.05	-.05	-.02	.06
P3_c1	.09	**.01**	.09	.12	-.09	.03	.03
P3_c32	**.00**	-.13	-.08	.02	-.09	-.03	.04
P3_c26r	.06	-.01	**.01**	.01	-.08	.01	.04
P3_c24r	.07	.00	.01	**.15***	-.10	-.03	.04

Anmerkungen. N = 199. Aufbau Parcelname: P1 bis P3 = Aufgabennummer, c = Gewissenhaftigkeit, zweite Zahl = Nummer des IPIP-Items der Kontingenz, r = Zielperson besitzt geringe Ausprägung auf dem Faktor Gewissenhaftigkeit. Die fett gedruckten Korrelationen markieren die Ergebnisse, wenn PU-Parcels und AB5C-Skala sich auf dasselbe Item beziehen bzw. dasselbe Item enthalten. Für eine bessere Übersicht über die Ergebnisse werden keine Konfidenzintervalle der Korrelationen berichtet.

*p < .05. **p < .01.

4.3.6.3 Zusammenhang PU und Big Five. Abschließend wurden die Zusammenhänge zwischen PU und den Big Five-Faktoren auf latenter Ebene geschätzt. Indikatoren für PU stellten die sechs Aufgaben-Parcels dar und als Indikatoren für die Big Five wurden die gebildeten Adjektiv-Parcels (alle drei Parcel-Sets) verwendet. Die Ergebnisse unter Verwendung des Parcel-Sets 1 sind Tabelle 4.12 und Tabelle 4.13 zu entnehmen (Modell-Fit unter Verwendung des MLM-Schätzers: χ^2 = 262.24, df = 174, p < .001, CFI = .95, RMSEA = .05, 90% CI = [.04, .06]). Bei der Interpretation der Faktorkorrelationen in Tabelle 4.13 ist erneut zu beachten, dass beim standardisierten Distanzscoring höhere Werte eine schlechtere Leistung anzeigen. Die Ergebnisse unter Verwendung von Parcel-Set 2 und Parcel-Set 3 finden sich in den Tabellen A8 und A9 in Anhang A.

Tabelle 4.12

Ergebnis der konfirmatorischen Faktorenanalyse der Personality Understanding (PU)-Parcels (standardisiertes Distanzscoring) und Adjektiv-Parcels (Parcel-Set 1)

	standardisierte Faktorladung					
Parcel	PU	O	C	E	A	N
PU1n	.75***					
PU1o	.54***					
PU2n	.69***					
PU2o	.69***					
PU3n	.90***					
PU3o	.83***					
oa_p1_o1		.70***				
oa_p1_o2		.76***				
oa_p1_o3		.83***				
oa_p1_c1			.72***			
oa_p1_c2			.86***			
oa_p1_c3			.86***			
oa_p1_e1				.72***		
oa_p1_e2				.81***		
oa_p1_e3				.94***		
oa_p1_a1					.79***	
oa_p1_a2					.75***	
oa_p1_a3					.87***	
oa_p1_n1						.75***
oa_p1_n2						.56***
oa_p1_n3						.84***

Anmerkungen. $N = 200$. O = Offenheit für Erfahrungen (bzw. Intellect bei Ostendorf, 1990); C = Gewissenhaftigkeit; E = Extraversion (bzw. Surgency bei Ostendorf, 1990); A = Verträglichkeit; N = Neurotizismus. Faktorkorrelationen siehe Tabelle 4.13.

***$p < .001$.

Tabelle 4.13

Faktorkorrelationen zwischen Personality Understanding (PU) und den Big Five-Faktoren

	PU	O	C	E	A
O	-.13 [-.26, .00]				
C	.12 [-.02, .27]	.33 [.18, .48]			
E	-.09 [-.26, .09]	.23 [.08, .39]	.21 [.05, .36]		
A	-.11 [-.25, .03]	.06 [-.09, .21]	.19 [.05, .33]	.13 [-.01, .28]	
N	-.03 [-.16, .09]	-.32 [-.45, -.20]	-.08 [-.23, .07]	-.15 [-.33, -.04]	.11 [-.05, .28]

Anmerkungen. $N = 200$. O = Offenheit für Erfahrungen (bzw. Intellect bei Ostendorf, 1990); C = Gewissenhaftigkeit; E = Extraversion (bzw. Surgency bei Ostendorf, 1990); A = Verträglichkeit; N = Neurotizismus. Für die verwendeten Indikatoren siehe Tabelle 4.12. In eckigen Klammern sind die 95%-Konfidenzintervalle der Korrelationen dargestellt.

Unter Verwendung des dichotomen Scorings ergaben sich vergleichbare Ergebnisse, die in den Tabellen B8 bis B10 in Anhang B zu finden sind. Die Ergebnisse unter Verwendung der verschiedenen Parcel-Sets fallen hier ebenfalls vergleichbar aus.

Insgesamt zeigten sich für die Zusammenhänge zwischen PU und den Big Five über beide Scoring-Methoden hinweg maximal kleine Korrelationen, wobei der Modell-Fit nicht bei allen Parcel-Sets gut ausfiel. Korrelationen $\geq |.10|$ fanden sich über alle geschätzten Modelle beider Scoring-Methoden hinweg für die Big Five-Faktoren Offenheit für Erfahrungen, Gewissenhaftigkeit und Verträglichkeit, wobei bei Offenheit und Verträglichkeit eine höhere Ausprägung auf der Persönlichkeitseigenschaft mit einer besseren Leistung bei PU einherging und bei Gewissenhaftigkeit sich das umgekehrte Muster zeigte.

4.4 Diskussion

Im Rahmen der ersten Studie wurden AcquA-Aufgaben zur Erfassung von PU konstruiert, für die sich zufriedenstellende Ergebnisse hinsichtlich faktorieller Validität und Reliabilität ergaben. Entsprechend den Annahmen konnten für die Konstrukte PU und EU zwei hoch korrelierte, aber separierbare Faktoren bestätigt werden. Während die Zusammenhänge zwischen PU und den Big Five-Persönlichkeitsfaktoren maximal klein und damit weitestgehend erwartungskonform ausfielen, konnte die Annahme, dass ein höherer Zusammenhang zwischen PU und Persönlichkeit der Testperson vorliegt, wenn Ziel- und

Testperson eine ähnliche Ausprägung auf der betrachteten Persönlichkeitseigenschaft aufweisen, nicht bestätigt werden.

Die Eindimensionalität der AcquA-PU Aufgaben konnte unter Verwendung beider Scoring-Methoden überwiegend bestätigt werden, wobei der RMSEA beim standardisierten Distanzscoring nicht ganz das Kriterium von ≤ .05 für einen guten Fit (Browne & Cudeck, 1992) erreichte. Während die Reliabilitätsschätzungen auf Basis des standardisierten Distanzscorings mit Werten von .85 für α und ω als gut beurteilt werden können, fielen die Ergebnisse unter Verwendung des dichotomen Scorings deutlich schlechter und mit lediglich .62 für α und ω nicht zufriedenstellend aus. Der Wert für Cronbachs α blieb beim dichotomen Scoring somit unterhalb der Ergebnisse von Hellwig et al. (2020), die unter Verwendung des dichotomen Scorings sowie ebenfalls sechs Aufgaben-Parcels Werte von .71 und .72 erzielten. Zwar zeigte sich auch in anderen Studien aus dem Bereich der SJTs (De Leng et al., 2017) sowie unter Verwendung von AcquA-EU Aufgaben (Pisters & Schulze, 2017), dass die Reliabilität unter Verwendung der standardisierten Distanzen höher ausfällt als beim dichotomen Scoring, allerdings war dort die Diskrepanz – wenn überhaupt vorhanden – deutlich geringer. In Übereinstimmung mit der geringen Reliabilität fielen auch die standardisierten Faktorladungen im Messmodell für PU sowie die Trennschärfen der selektierten Items beim dichotomen Scoring teilweise zu gering aus. Bei der Interpretation der faktoriellen Validität sowie der Reliabilität der AcquA-PU Aufgaben muss zudem einschränkend berücksichtigt werden, dass inhaltlich sehr homogene Aufgaben konstruiert wurden. Den Kontingenzen aller sechs Aufgaben wurden Items der Gewissenhaftigkeit zugrunde gelegt und zudem basieren die Kontingenzen von jeweils zwei Aufgaben auf demselben Item. Die daraus resultierende inhaltliche Homogenität kann möglicherweise zu einer Überschätzung der Eindimensionalität sowie der Reliabilität geführt haben. Daher ist für zukünftige Studien die Konstruktion weiterer Aufgaben, die einem anderen Big Five-Faktor zugeordnet werden können, erforderlich.

Darüber hinaus muss hervorgehoben werden, dass die AcquA-PU Aufgaben extrem leicht waren. Dies wird insbesondere beim dichotomen Scoring bei Betrachtung der dort bestimmten Mittelwerte der PU-Parcels, des Gesamtscores sowie der selektierten Items deutlich. Letztere nahmen Werte bis inklusive .99 an (24 Items zeigten einen Itemmittelwert > .95; vgl. Tabelle A1 in Anhang A), sodass Items beibehalten wurden, die kaum Information liefern. Aber auch beim standardisierten Distanzscoring deutet die Schiefe der Parcels und des Gesamtscores darauf hin, dass die Aufgaben zu leicht waren. Einerseits ist die hohe Leichtigkeit der Aufgaben damit zu vereinbaren, dass in der Vergangenheit nur geringe

interindividuelle Unterschiede in der akkuraten Beurteilung der Persönlichkeit anderer Personen gefunden wurden (Biesanz, 2010; Kenny, 1994; Funder, 1999). Rogers und Biesanz (2019) konnten andererseits zeigen, dass hierbei die Zielperson eine wichtige Rolle spielt. Ihren Ergebnissen zufolge lassen sich interindividuelle Unterschiede zwischen Beurteiler:innen vor allem dann finden, wenn gute Zielpersonen beurteilt werden, also solche, über die ausreichend relevante Hinweise zur Verfügung stehen. Für die AcquA-PU Aufgaben kann auf Grund der Konstruktionsweise, die die Präsentation der relevanten Information in der Aneignungsphase sicherstellt, angenommen werden, dass ausschließlich gute Zielpersonen beurteilt werden müssen. In zukünftigen Studien sollte die extreme Leichtigkeit daher weiter untersucht und die Schwierigkeit der Aufgaben erhöht werden. Zum einen sollte die Schwierigkeit, soweit möglich, in der Itemselektion berücksichtigt werden und zum anderen sollten gezielt schwierige Aufgaben konstruiert werden, was im Bereich EU ebenfalls bereits angestrebt wurde (Hellwig, 2016). Die extreme Leichtigkeit und die damit einhergehenden schiefen Verteilungen der Items und Aufgaben stellen zudem ein Problem für die darauf aufbauenden statistischen Analysen dar. Beispielsweise konnten Curran et al. (1996) zeigen, dass bei steigender Abweichung von einer multivariaten Normalverteilung selbst robuste Teststatistiken wie der Satorra-Bentler χ^2 eine vergleichsweise geringere Power zur Identifikation von falsch spezifizierten Modellen aufweisen. Darüber hinaus gehen auch diagnostische Schwierigkeiten mit den sehr leichten Items und Aufgaben einher, da so keine Differenzierung von Personen des gesamten Fähigkeitsbereichs möglich ist.

Die Umsetzung des neuen Formats der Aufgabendarstellung kann auf Basis der Ergebnisse zum Messmodell für PU (vgl. auch Messmodell für EU) als erfolgreich beurteilt werden. Weder im Messmodell noch im Hinblick auf die deskriptiven Statistiken zeigten die Aufgaben im 3D-Format schlechtere Ergebnisse als die Aufgaben im 2D-Format – allerdings auch keine besseren. Da die Aufgabenkonstruktion im 3D-Format auf Grund der Komplexität von iClone 6 (Reallusion, 2014) deutlich aufwändiger und zeitintensiver ist, könnte die Frage entstehen, warum das 3D-Format bevorzugt werden sollte, wenn es keine psychometrischen Vorteile mit sich bringt. Wie Hellwig (2016) bereits anmerkte, sind die Aufgaben im 2D-Format sehr textlastig und beanspruchen somit die verbalen Fähigkeiten der Testpersonen. Durch das neue 3D-Format kann die Textlastigkeit deutlich reduziert werden, vor allem wenn die Aufgaben zusätzlich vertont werden, was ein Ziel zukünftiger Konstruktionen sein sollte. Auf diese Weise können auch die Akzeptanz der Testpersonen erhöht und die ökologische Validität gesteigert werden. Zudem sollte beachtet werden, dass die Durchführungszeit für die AcquA-Aufgaben relativ hoch ist und in der vorliegenden Studie bei etwa einer Stunde für

sechs Aufgaben lag. Weniger textlastige Aufgaben könnten in dem Zusammenhang auch dazu beitragen, dass die Motivation der Testpersonen aufrecht erhalten bleibt. Ein wichtiger Unterschied zwischen beiden Formaten ist zudem, dass beim 3D-Format die beobachtbaren Verhaltensweisen zunächst einmal auf die intendierte Weise verstanden werden müssen. Beispielsweise müssen ein Kopfnicken auch als zustimmendes Verhalten oder das Liegenlassen eines Gegenstandes als unordentliches Verhalten interpretiert werden. Dies kann dem letzten Schritt des RAM, in dem es um die valide Nutzung erkannter Hinweise geht (Funder, 1995), zugeordnet und somit als Teil von PU angesehen werden (Erkennen der Regelhaftigkeiten). Das 3D-Format stellt somit eine geeignetere Operationalisierung von PU dar, vorausgesetzt, die Verhaltensweisen werden korrekt modelliert, sodass eine richtige Interpretation grundsätzlich möglich ist.

Durch die Untersuchung des latenten Zusammenhangs zwischen den AcquA-PU und AcquA-EU Aufgaben konnte bestätigt werden, dass sich das AcquA-Testdesign entsprechend der Annahmen von Schulze und Roberts (2015) neben EU auch zur Operationalisierung weiterer und konzeptuell sehr ähnlicher Fähigkeitskonstrukte eignet. Zudem konnte hierdurch erste konvergente Validitätsevidenz für die AcquA-PU Aufgaben gesammelt werden. Die sehr hohe latente Korrelation spiegelt die inhaltlichen Überschneidungen der beiden Konstrukte sowie die stark überlappende Operationalisierung angemessen wider. Sie kann daher auch nicht als zu hoch angesehen werden, zumal das zweifaktorielle Modell einen signifikant besseren Fit zeigte als das einfaktorielle Modell. Darüber hinaus ist der latente Zusammenhang zwischen PU und EU mit einem Wert von .66 (standardisiertes Distanzscoring) beziehungsweise .74 (dichotomes Scoring) nicht viel höher als die Zusammenhänge von logischem Schlussfolgern in den klassischen Inhaltsbereichen. Liepmann et al. (2007) betrachteten im Rahmen von Untersuchungen zur Validität des Intelligenz-Struktur-Tests 2000 R auf Basis einer exploratorischen Faktorenanalyse auch die latenten Zusammenhänge zwischen verbalem, numerischem und figuralem schlussfolgernden Denken. Hier resultierten Korrelationen von .41 und .50 (verbal und figural), .54 und .55 (numerisch und figural) sowie .59 und .67 (numerisch und verbal; Liepmann et al., 2006).

Die Ergebnisse zum CTC(M-1)-Modell, das bei beiden Scoring-Methoden keinen besseren Fit aufwies als ein zweifaktorielles Modell ohne Methodenfaktor, unterstützen die Schlussfolgerung, dass die Umsetzung des neuen 3D-Formats erfolgreich war. Beim standardisierten Distanzscoring zeigten sich keine systematischen und sinnvoll interpretierbaren Ladungen der Aufgaben im 3D-Format auf dem Methodenfaktor. Beim dichotomen Scoring zeigen sich zwar durchweg positive standardisierte Ladungen, allerdings

waren diese mit ≤ .25 nicht sehr groß. Zudem muss beachtet werden, dass beim dichotomen Scoring kein zufriedenstellender Modellfit resultierte und das CTC(M-1)-Modell erst nach dem Setzen von Restriktionen schätzbar war. Insgesamt betrachtet scheint es den modellierten Methodenfaktor nicht zu geben, sodass angenommen werden kann, dass die beiden Formate keinen substantiell unterschiedlichen Einfluss auf die Leistung in den AcquA-Aufgaben haben.

Die latenten Zusammenhänge zwischen PU und den Big Five fielen über alle geschätzten Modelle beider Scoring-Methoden hinweg maximal klein und fast ausschließlich nicht signifikant aus. Insgesamt können die Ergebnisse daher als Evidenz für diskriminante Validität der AcquA-PU Aufgaben interpretiert werden. Die deskriptiv etwas höheren, aber immer noch kleinen Zusammenhänge von PU zu Verträglichkeit sowie Offenheit für Erfahrungen passen jedoch eher zu Ergebnissen, die bei Betrachtung von persönlichkeitsrelevantem Wissen und entsprechenden Komponenten der Akkuratheit erzielt wurden. So resultierten für Verträglichkeit positive Beziehungen zur Normative Accuracy, nicht aber zur Distinctive Accuracy (Biesanz, 2010; Letzring, 2015). Zudem zeigten sich bei Letzring (2008) sowie Vogt und Colvin (2003) positive Beziehungen zwischen Verträglichkeit und der Profile Accuracy, die eine normative Komponente der Akkuratheit beinhaltet (Funder, 1999; Furr, 2008; Kenny & Winquist, 2001). Für Offenheit für Erfahrungen zeigte sich ein mittlerer Zusammenhang mit DI (Christiansen et al., 2005; siehe aber auch de Vries, et al., 2021), der auch für klassisches Wissen beziehungsweise Gc üblich ist (Ackerman & Heggestad, 1997). Biesanz (2010) konnte im Gegensatz dazu für Offenheit für Erfahrungen nur Zusammenhänge zur Distinctive Accuracy finden, nicht aber zur Normative Accuracy. Ob und gegebenenfalls was aus den etwas höheren Zusammenhängen zwischen PU und Verträglichkeit sowie Offenheit für Erfahrungen geschlossen werden kann, ist zum aktuellen Zeitpunkt allerdings unklar.

Bei Gewissenhaftigkeit ging eine höhere Ausprägung tendenziell mit einer schlechteren Leistung bei PU einher. Dies erscheint in Anbetracht dessen, dass bei allen PU-Aufgaben Schlussfolgerungen über gewissenhaftes Verhalten gezogen werden mussten und vor dem Hintergrund, dass Ähnlichkeit zwischen Testperson und Zielperson eventuell einen positiven Einfluss auf die Testleistung hat, zunächst einmal ungewöhnlich. Allerdings ist ein negativer Zusammenhang zur Gewissenhaftigkeit auch beim klassischen logischen Schlussfolgern sowie allgemeiner Intelligenz kein ungewöhnliches Ergebnis (Moutafi et al., 2004; Zajenkowski & Stolarski, 2015; siehe aber auch Rikoon, et al., 2016).

Insgesamt sollten die kleinen Korrelationen zwischen PU und den drei Big Five-Faktoren auf Grund der überwiegend fehlenden Signifikanz mit Vorsicht interpretiert werden.

Darüber hinaus muss hierbei beachtet werden, dass mit den Adjektiven von Ostendorf (1990) möglicherweise keine angemessene Operationalisierung der Big Five gewählt wurde. Es mussten im Rahmen der Itemselektion fast die Hälfte der Adjektive eliminiert werden, um eindimensionale Skalen zu erhalten. Ein möglicher Grund hierfür ist das Alter der Adjektive (Ostendorf, 1990). Die Sprache entwickelt sich fortlaufend, sodass einige Adjektive den Testpersonen vermutlich nicht mehr geläufig waren oder unterschiedlich verstanden wurden. Der Faktor Offenheit für Erfahrungen wurde bei Ostendorf (1990) zudem nicht als solcher konzeptualisiert, sondern als Intellekt, was eine alternative Interpretation des fünften Big Five-Faktors darstellt (Digman, 1990; McCrae & John, 1992; Ostendorf, 1990). Dies hat letztendlich dazu geführt, dass die selektierten Adjektive eher die selbsteingeschätzte Intelligenz der Testpersonen erfasst haben (vgl. Tabelle A4 in Anhang A). Der Zusammenhang zwischen PU und den Big Five sollte daher noch einmal mit einem etablierten Instrument repliziert werden.

Dass die Annahme hinsichtlich der Ähnlichkeit zwischen Testperson und Zielperson nicht bestätigt werden konnte, kann mehrere Ursachen haben. Zum einen ist denkbar, dass die hypothesengeleitet gebildeten PU-Parcels zu heterogen waren, um eine spezifische Facette beziehungsweise ein spezifisches Item der Gewissenhaftigkeit angemessen widerzuspiegeln. In der Anwendungsphase der AcquA-Aufgaben müssen Verhaltensweisen eingeschätzt werden, die der Kontingenz aus der Aneignungsphase entsprechen (likely-Items) und solche, die der Kontingenz nicht entsprechen (unlikely-Items; Schulze & Roberts, 2015). In der Regel müssen deutlich mehr unlikely-Items eingeschätzt werden (ca. vier bis fünf von insgesamt sechs Items bei den Aufgaben der vorliegenden Studie), sodass die Parcels hauptsächlich aus diesen Items bestanden. Die unlikely-Items der AcquA-PU Aufgaben können wiederum sehr heterogenes Verhalten beinhalten und neben einer anderen Ausprägung der Gewissenhaftigkeit auch eine andere Persönlichkeitseigenschaft widerspiegeln (vgl. Abschnitt 4.2.1.5). Möglicherweise hat diese Heterogenität dazu geführt, dass die Ähnlichkeit bei allen in den unlikely-Items modellierten Persönlichkeitseigenschaften eine Rolle bei der Aufgabenbearbeitung gespielt hat. Hinzu kommt, dass die Reliabilität der hypothesengeleitet gebildeten PU-Parcels größtenteils eher gering ausfiel. Zudem muss auch hier auf die extreme Schiefe und Leichtigkeit der Parcels hingewiesen werden, was zu einer Unterschätzung der Korrelationen mit den AB5C-Skalen geführt haben könnte.

Darüber hinaus ist kritisch zu sehen, dass die Ähnlichkeit zwischen Testperson und Zielperson primär auf Basis eines einzelnen Items spezifiziert wurde. Studien, die die Ähnlichkeit in der Persönlichkeit zwischen Beurteiler:in und Zielperson untersucht haben, haben die Persönlichkeit in der Regel deutlich breiter erfasst (z.B. Funder et al., 1995; Kurtz &

Sherker, 2003; Vogt & Colvin, 2003). Eine mögliche Verbesserung für zukünftige Untersuchungen zur Rolle der Ähnlichkeit bei den AcquA-PU Aufgaben wäre daher, die Persönlichkeitseigenschaften, für die die Ähnlichkeit betrachtet werden soll, umfangreicher und auf Basis mehrerer Items in den Kontingenzen der Zielperson abzubilden.

Eine weitere potenzielle Fehlerquelle stellen die neu gebildeten AB5C-Skalen dar, bei denen keine stabile Zuordnung der IPIP-Items vorgenommen werden konnte. Einerseits weisen die AB5C-Skalen eine gewisse inhaltliche Homogenität auf: Skala III+/I+ vs. III-/I- beinhaltet Items mit Fokus auf Ordnung, Skala III+/IV+ vs. III-/IV- Items mit Fokus auf Prokrastination oder Skala I+/I+ vs. I-/I- Items mit Fokus auf Kommunikation mit anderen Personen. Andererseits sind durch den AB5C-Algorithmus inhaltlich fast redundante Items unterschiedlichen Skalen zugeordnet worden: „Ich plane gerne im Voraus." (III+/I+ vs. III-/I-) und „Ich plane nicht im Voraus." (III+/V+ vs. III-/V-) sowie „Ich finde es schwierig, mit der Arbeit anzufangen." (III+/IV+ vs. III-/IV-) und „Ich habe Schwierigkeiten, Aufgaben zu beginnen" (III+/V+ vs. III-/V-). Dieses Ergebnis spiegelt ein mögliches Problem des AB5C-Modells wider, das bereits von Saucier und Ostendorf (1999) diskutiert wurde. Die Autoren vermuten, dass für eine stabile Zuordnung vier- oder fünfstellige Stichprobengrößen benötigt werden. Hinzu kommt, dass die Zuordnung der Items zu den AB5C-Skalen von den verwendeten Marker-Adjektiven abhängig ist, die in der vorliegenden Studie ebenfalls nicht frei von Problemen sind (vgl. oben). Insgesamt betrachtet ist es daher auch nicht überraschend, dass die auf Basis der übersetzten IPIP-Items gebildeten AB5C-Skalen keine Replikation der Skalen von Goldberg (1999) darstellen.

Letztendlich muss auch diskutiert werden, dass in der vorliegenden Studie der Fokus auf die tatsächliche Ähnlichkeit gelegt wurde, da eine Berücksichtigung der angenommenen Ähnlichkeit nicht ohne Weiteres möglich war. Hierfür hätten die Testpersonen einschätzen müssen, wie sie selbst in den jeweiligen Anwendungsphasen der AcquA-PU Aufgaben reagiert hätten. Das heißt, die Testpersonen hätten für die einzuschätzenden Reaktionen der Zielpersonen zusätzlich einen Selbstbericht abgeben müssen. Es bleibt in der vorliegenden Studie somit unklar, ob die Testpersonen bei eventuell vorliegender tatsächlicher Ähnlichkeit auch Ähnlichkeit zur Zielperson angenommen haben. Ein Effekt auf die Testleistung ist aber nur dann zu erwarten, wenn beides vorliegt (Paunonen & Hong, 2013; Vogt & Colvin, 2003). Nach den Ergebnissen einer Meta-Analyse von Thielmann et al. (2020) nehmen Beurteiler:innen insbesondere für die Persönlichkeitsfaktoren Offenheit für Erfahrungen und Verträglichkeit beziehungsweise Ehrlichkeit-Bescheidenheit, aber unter anderem nicht für die Gewissenhaftigkeit, Ähnlichkeit zur Zielperson an. Dies könnte erklären, warum in der

vorliegenden Studie, in der der Fokus auf der Gewissenhaftigkeit lag, kein Effekt gefunden wurde. Darüber hinaus ist es auch möglich, dass ein Zusammenhang zwischen Ähnlichkeit und Akkuratheit in Übereinstimmung mit Human und Biesanz (2011, 2012) sowie Letzring (2015) nur für die normative Komponente der Akkuratheit zu finden ist und durch den Fokus von PU und dem AcquA-Testdesign auf logisches Schlussfolgern hier keine Rolle gespielt hat. Alles in allem ist auf Grund der vielfältigen Kritikpunkte keine eindeutige Interpretation der vorliegenden Ergebnisse möglich, sodass die Rolle der Ähnlichkeit zwischen Testperson und Zielperson bei den AcquA-PU Aufgaben weiterhin als unklar angesehen werden muss.

Zusammenfassend betrachtet konnte die erste Studie der vorliegenden Arbeit erste Hinweise auf faktorielle, konvergente und diskriminante Validität der neuen AcquA-Aufgaben zur Erfassung von PU liefern. Insbesondere die Untersuchung des Zusammenhangs zwischen PU und EU kann als besonders kritischer Test für PU sowie das AcquA-Testdesign angesehen werden. Die Ergebnisse müssen allerdings durch weitere Evidenz ergänzt werden, um die Interpretation zuzulassen, dass mit den AcquA-PU Aufgaben logisches Schlussfolgern im Inhaltsbereich Persönlichkeit erfasst wird. Zudem sollte in einem nächsten Schritt eine Erweiterung des Pools der AcquA-PU Aufgaben im 3D-Format vorgenommen werden – zum einen, da derzeit nur drei Aufgaben in diesem Format vorliegen und zum anderen, da es sich hierbei ausschließlich um Aufgaben mit Fokus auf die Gewissenhaftigkeit handelt. Mittelfristig ist auf Grund der aktuell zu leichten Aufgaben auch eine gezielte Konstruktion schwieriger Aufgaben zwingend erforderlich. Mit den im Rahmen der ersten Studie konstruierten Aufgaben wurde für solche und weitere Studien allerdings eine gute Ausgangsposition geschaffen.

5. Studie 2

5.1 Ziele

In der vorherigen Studie konnte gezeigt werden, dass das AcquA-Testdesign neben EU auch zur Operationalisierung weiterer und konzeptuell sehr ähnlicher Konstrukte wie PU eingesetzt werden kann. Des Weiteren konnte ein neues 3D-Format der AcquA-Aufgaben, für das kein substantieller Methodeneffekt gefunden wurde, erfolgreich umgesetzt werden. Daher wurde in der zweiten Studie der Fokus auf die AcquA-PU Aufgaben im neuen 3D-Format gelegt. Daran anschließend bestand das erste Ziel in der Überarbeitung der drei bereits vorhandenen Aufgaben sowie in der Ergänzung neuer Aufgaben mit Fokus auf einen anderen Persönlichkeitsfaktor. Die Überarbeitung der vorhandenen Aufgaben umfasste hauptsächlich die Vertonung der Dialoge. Einerseits sollte so die Bearbeitungszeit der Aufgaben reduziert werden und andererseits sollte auf diese Weise kontrolliert werden, dass die Testpersonen ihre Aufmerksamkeit nicht zu stark auf die mittels Sprechblasen präsentierten Dialoge lenken und dabei relevante Verhaltensweisen der Zielpersonen übersehen. Auf Grund der Überarbeitungen und Ergänzungen der Aufgaben wurde im Rahmen des ersten Ziels auch eine erneute Überprüfung der psychometrischen Qualität, inklusive Überprüfung der angenommenen Eindimensionalität sowie Schätzung der Reliabilität, vorgenommen.

Das zweite Ziel bestand in der Sammlung weiterer konvergenter Validitätsevidenz. Dadurch, dass PU die Fähigkeit zum logischen Schlussfolgern im Inhaltsbereich Persönlichkeit darstellen soll, ist die Untersuchung des Zusammenhangs zum logischen Schlussfolgern in den klassischen Inhaltsbereichen unerlässlich. Geht man davon aus, dass bei PU dieselbe kognitive Operation beteiligt ist wie beim verbalen, numerischen und figuralen logischen Schlussfolgern und sich hiervon nur durch den anderen Inhaltsbereich unterscheidet (vgl. Abschnitt 3.1.2), so ist zwischen PU und logischem Schlussfolgern auf latenter Ebene ein großer positiver Zusammenhang zu erwarten. Wie in Abschnitt 3.1.2 ebenfalls berichtet wurde, ist die vorhandene empirische Evidenz hinsichtlich eines Zusammenhangs zu Intelligenz oder dem logischen Schlussfolgern aus den Bereichen akkurater Persönlichkeitsbeurteilungen, SU und DI zwar sehr heterogen, aber auch nur mit Einschränkungen auf den Bereich PU übertragbar. Der angenommene große Zusammenhang zwischen PU und logischem Schlussfolgern basiert daher in erster Linie auf theoretischen Annahmen zum Konstrukt PU. Unter Berücksichtigung der konkreten Operationalisierung durch das AcquA-Testdesign liefern zudem die Ergebnisse von Hellwig et al. (2020) sowie Schulze und Jobmann (2016) Unterstützung für den

angenommen Zusammenhang. Hellwig et al. (2020) ermittelten zwischen AcquA-EU und einer Aufgabe zur Erfassung von figuralem logischen Schlussfolgern auf latenter Ebene eine Korrelation von .46. Bei Schulze und Jobmann (2016) resultierte eine latente Korrelation von .54 zwischen AcquA-EU und Aufgaben zur Erfassung von Arbeitsgedächtniskapazität, die eine große Überlappung mit logischem Schlussfolgern aufweist (Süß et al., 2002).

Das dritte Ziel ergab sich aus den Ergebnissen der vorherigen Studie und bezog sich erneut auf den Zusammenhang zwischen PU und der Persönlichkeit der Testperson. Zum einen sollte zur Sicherstellung diskriminanter Validitätsevidenz der maximal geringe Zusammenhang zwischen PU und den Big Five repliziert werden. In Studie 1 wurde dieser unter Verwendung der Adjektive von Ostendorf (1990) geschätzt, die zuvor auf Grund fehlender Eindimensionalität der Skalen um fast die Hälfte reduziert werden mussten. Es ist somit fraglich, ob die selektierten Adjektive die Faktoren angemessen abgebildet haben. Der Zusammenhang sollte daher unter Verwendung eines etablierten Instruments zur Erfassung der Big Five erneut untersucht werden. Zum anderen konnte in Studie 1 kein höherer Zusammenhang zwischen PU und Persönlichkeit der Testperson gefunden werden, wenn Testperson und Zielperson eine ähnliche Ausprägung hinsichtlich der betrachteten Persönlichkeitseigenschaft aufweisen. Wie bereits in Abschnitt 4.4 angemerkt wurde, könnte dies unter anderem an den verwendeten AB5C-Skalen gelegen haben. So konnte weder für den Faktor Gewissenhaftigkeit noch für den Faktor Extraversion eine eindeutige und stabile Zuordnung der IPIP-Items zu den AB5C-Skalen erzielt werden. In Studie 2 wurde daher überprüft, ob sich die Zuordnung der IPIP-Items zu den AB5C-Skalen der Gewissenhaftigkeit auf Basis einer neuen Stichprobe replizieren lässt. Im Falle einer nicht erfolgreichen Replikation wurde eine alternative faktorenanalytische Skalenbildung geplant und in jedem Fall erneut exploriert, ob ein höherer Zusammenhang zwischen PU und Persönlichkeit der Testperson resultiert, wenn Testperson und Zielperson eine ähnliche Ausprägung hinsichtlich der betrachteten Persönlichkeitseigenschaft aufweisen.

5.2 Methode

5.2.1 Stichprobe

Insgesamt nahmen 218 Personen an der Studie teil. Eine Person brach die Erhebung sehr früh ab, sodass keine verwertbaren Daten erhoben werden konnten. Zudem erfüllten sieben Personen das auf Grund des hohen verbalen Anteils festgelegte Sprachkriterium (≥ 10 Jahre Deutschkenntnisse; vgl. Studie 1) nicht und wurden ebenfalls von den Analysen

ausgeschlossen. Darüber hinaus wurden sechs weitere Personen ausgeschlossen, die bereits an der vorherigen Studie teilgenommen hatten. Zwar wurde das Testmaterial von Studie 1 nach Studie 2 überarbeitet und erweitert, allerdings war die Schnittmenge bei den für die Studie zentralen AcquA-PU Aufgaben immer noch so groß, dass Erinnerungs- und Übungseffekte nicht ausgeschlossen werden konnten.

Die vollständige Analysestichprobe bestand somit aus $N = 204$ Personen (134 weiblich, 69 männlich, 1 divers), die über Aushänge an der Bergischen Universität Wuppertal, Social Media sowie im privaten Umfeld der Versuchsleiterinnen rekrutiert wurden. Das mittlere Alter der Stichprobe lag bei 26.76 Jahren ($SD = 10.08$; Range: 17 bis 68). Insgesamt 191 Personen gaben Deutsch als Muttersprache an und die übrigen 13 Personen sprachen seit mindestens 10 Jahren Deutsch. Die häufigste Angabe bei der Frage nach dem höchsten akademischen oder schulischen Abschluss war die Fachhochschul- oder Hochschulreife/Abitur ($n = 113$), danach folgten Bachelor ($n = 59$), Realschul- oder gleichwertiger Abschluss ($n = 13$), Master/Diplom ($n = 12$), noch in schulischer Ausbildung ($n = 3$), Haupt-(Volks-)schulabschluss ($n = 2$) und Promotion ($n = 2$). Zudem gaben 151 Personen der Analysestichprobe an, dass Sie derzeit studieren. Die Psychologiestudierenden waren hier die größte Gruppe ($n = 57$).

5.2.2 Messinstrumente

5.2.2.1 Personality Understanding. Zur Erfassung von PU wurden insgesamt sechs AcquA-PU Aufgaben eingesetzt. Zum einen wurden die drei Aufgaben PU1, PU2 und PU3 (3D-Format) in überarbeiteter Form aus der vorherigen Studie übernommen. Die größte Veränderung bestand in der Vertonung der Dialoge zwischen den interagierenden Avataren. Hierbei sollten Emotionen als mögliche, nicht intendierte Einflussquelle soweit wie möglich ausgeschlossen werden. Da Emotionen auch durch die Sprache vermittelt werden (Banse & Scherer, 1996; Laukka et al., 2016; Scherer, 2003), wurden die Personen, die die Dialoge eingesprochen haben, instruiert, dies möglichst neutral und ohne den Ausdruck einer bestimmten Emotion zu tun. Die Instruktion, die in Anlehnung an Laukka et al. (2016) erstellt wurde, enthielt zudem die Anweisung, den Text deutlich und in einem moderaten Tempo vorzulesen. Zudem wurden einige aus der Vertonung resultierende kleinere Änderungen an den 3D-Simulationen vorgenommen: Da die Sprechblasen in der vorherigen Aufgabenversion deutlich länger eingeblendet wurden, als der vertonte Text dauerte, wurden Zwischensequenzen, in denen nichts passiert, herausgeschnitten oder gekürzt. Zudem mussten Dauer und Latenz einiger Bewegungen der Avatare an den gesprochenen Text angepasst

werden. Darüber hinaus gab es unabhängig von der Vertonung kleinere generelle Überarbeitungen der Aufgaben (z.B. leichte Korrekturen der Bewegungen). Inhaltlich wurden die Aufgaben und insbesondere die relevanten Kontingenzen nicht verändert.

Die drei überarbeiteten Aufgaben wurden durch die drei neu konstruierten Aufgaben PU4, PU5 und PU6 ergänzt.[10] Als Grundlage für die relevanten Kontingenzen dieser Aufgaben wurden IPIP-Items folgender drei AB5C-Facetten der Verträglichkeit ausgewählt (Goldberg, 1999): Understanding (II+/II+ vs. II-/II-), Pleasantness (II+/IV+ vs. II-/IV-), Cooperation (II+/I- vs. II-/I+). Die Konstruktion erfolgte nach den in Studie 1 beschriebenen Prinzipien mit dem Unterschied, dass die Dialoge direkt vertont wurden. Eine Übersicht über verschiedene Aspekte der sechs eingesetzten AcquA-PU Aufgaben ist in Tabelle 5.1 zu finden.

Die Auswertung der AcquA-PU Aufgaben erfolgte wie in Studie 1 primär auf Basis des standardisierten Distanzscorings. Das dichotome Scoring wurde zusätzlich verwendet, um die Ergebnisse des standardisierten Distanzscorings zu ergänzen.

Tabelle 5.1

Übersicht über verschiedene Aspekte der in Studie 2 eingesetzten AcquA-Aufgaben zur Erfassung von Personality Understanding (PU)

		Anzahl		
Aufgabe[a]	AB5C-Facette der Kontingenzen	Zielpersonen in AP1	relevante Kontingenzen in AP1	Einschätzungen in AP2
PU1	Efficiency	2	2	12
PU2	Orderliness	2	2	12
PU3	Conscientiousness	2	4	24
PU4	Understanding	2	2	12
PU5	Pleasantness	2	2	12
PU6	Cooperation	2	2	12
Gesamt		12	14	84

Anmerkungen. AP1 = Aneignungsphase; AP2 = Anwendungsphase.

[a] Alle Aufgaben wurden in der Version 2 eingesetzt, d.h. die Dialoge wurden ohne den Ausdruck einer bestimmten Emotion vertont.

[10] Diese Aufgaben wurden unter enger Betreuung im Rahmen der Masterarbeiten von Melanie Nauditt, Ronja Nippert und Rebekka Prielipp konstruiert, bei denen ich mich an dieser Stelle ganz herzlich für die Arbeit und den Einsatz im Rahmen der Erhebungen bedanken möchte.

5.2.2.2 Logisches Schlussfolgern. Zur Erfassung der Fähigkeit zum logischen Schlussfolgern wurde Modul 4 aus dem Wilde Intelligenz Test 2 (WIT-2; Kersting et al., 2008) eingesetzt. Theoretische Grundlage des WIT-2 ist eine Modifikation von Thurstones (1938) Modell mehrerer gemeinsamer Faktoren, das sogenannte Modifizierte Modell der Primary Mental Abilities (MMPMA), das das ursprüngliche Modell um Annahmen hinsichtlich einer Facettenstruktur, Hierarchieebenen und Zusammenhängen zur Arbeitsgedächtniskapazität ergänzt (Kersting et al., 2008). Thurstones (1938) Modell umfasst unter anderem eine Fähigkeit zum logischen Schlussfolgern, die im WIT-2 durch Modul 4 unter der Bezeichnung s*chlussfolgerndes Denken* erfasst wird. Auf Basis der im MMPMA ergänzten Facetten-Annahme wird das schlussfolgernde Denken im WIT-2 analog zum BIS-Modell als kognitive Operation angesehen, die sich in unterschiedlichen Inhaltsbereichen zeigen kann. Operationalisiert wird das schlussfolgernde Denken daher mit Hilfe von verbalem, numerischem und figuralem Aufgabenmaterial (Kersting et al., 2008). Modul 4 wurde daher, und auf Grund der im Manual dokumentierten sehr guten psychometrischen Eigenschaften, als geeignet angesehen, um den Zusammenhang zwischen PU und logischem Schlussfolgern in den klassischen Inhaltsbereichen zu untersuchen.

Modul 4 des WIT-2 besteht aus den drei Untertests Analogien (verbales Material), Zahlenreihen (numerisches Material) sowie Abwicklungen (figurales Material), die jeweils 20 Items beinhalten (Kersting et al., 2008). Beim Untertest Analogien werden zwei Wörter vorgegeben, zwischen denen eine bestimmte Beziehung besteht, sowie ein drittes Wort. Aus fünf weiteren Wörtern soll dasjenige Wort ausgewählt werden, das dieselbe Beziehung zu dem dritten Wort aufweist, wie die ersten beiden Wörter untereinander zeigen. Der Untertest Zahlenreihen umfasst eben solche, die nach einer bestimmten Regel erstellt wurden. Diese Regel muss erkannt werden und die Zahlenreihen um die nächste Zahl ergänzt werden. Der Untertest Abwicklungen besteht aus Abbildungen von auseinandergefalteten geometrischen Körpern. Aus fünf zusammengefalteten Körpern soll jeweils derjenige ausgewählt werden, der aus der Faltvorlage erstellt werden kann. Bei den Untertests Analogien und Abwicklungen handelt es sich folglich um Multiple-Choice-Items. Der Untertest Zahlenreihen erfordert hingegen eine offene Antwort und alle Untertests werden mit einer festen Bearbeitungszeit durchgeführt. Laut Manual liegt die Dauer des gesamten Moduls bei etwa 35 Minuten und die Reliabilität bei $\alpha = .94$ (stratifiziertes Cronbachs α; Kersting et al., 2008).

5.2.2.3 Persönlichkeit der Testpersonen. Zur Erfassung der Big Five-Faktoren wurde die deutschsprachige Version des NEO-Fünf-Faktoren-Inventars nach Costa und McCrae

(NEO-FFI; Borkenau & Ostendorf, 2008) eingesetzt. Hierbei handelt es sich um eine Kurzversion der deutschsprachigen Version des NEO-PI-R (Ostendorf & Angleitner, 2004), mit der die fünf Faktoren Neurotizismus, Extraversion, Offenheit für Erfahrungen, Gewissenhaftigkeit sowie Verträglichkeit des FFM mit jeweils 12 Items erfasst werden können. Eine Operationalisierung auf Ebene von Facetten ist hingegen nicht möglich (Ostendorf & Angleitner, 2004). Das NEO-FFI ist ein Selbstberichtsverfahren, in dem die Testpersonen auf einer 5-stufigen Likert-Skala von *Starke Ablehnung* (0) bis *Starke Zustimmung* (5) angeben sollen, inwieweit die Aussagen auf sie selbst zutreffen. Als Reliabilitätsschätzungen werden im Manual von Borkenau und Ostendorf (2008) folgende Werte für Cronbachs α berichtet: .72 (Verträglichkeit), .75 (Offenheit für Erfahrungen), .81 (Extraversion), .84 (Gewissenhaftigkeit) und .87 (Neurotizismus).

Darüber hinaus wurden zur Untersuchung eines möglichen Effekts der Ähnlichkeit zwischen Testperson und Zielperson erneut die zur Konstruktion der AcquA-PU Aufgaben verwendeten AB5C-Facetten von Goldberg (1999) erhoben. Dies waren neben den bereits in Studie 1 eingesetzten und übersetzten Gewissenhaftigkeits-Facetten Efficiency, Orderliness und Conscientiousness auch die folgenden weiter oben genannten Facetten der Verträglichkeit (Goldberg, 1999): Understanding (10 Items, im Original α = .81), Pleasantness (12 Items, α = .76) sowie Cooperation (9 Items, α = .73 [für die Reliabilitätsschätzung wurden drei Items einer anderen Facette hinzugezogen; Goldberg, 1999]). Die Items der drei Verträglichkeits-Facetten wurden auf eine ähnliche Weise übersetzt wie die Items der Gewissenhaftigkeits-Facetten (vgl. Abschnitt 4.2.3.3). Um eine Skalenzuordnung nach dem AB5C-Algorithmus vornehmen zu können, wurden zudem erneut die 100 Marker-Adjektive von Ostendorf (1990) eingesetzt. Die IPIP-Items wie auch die Adjektive wurden von den Testpersonen auf einer 6-stufigen Likert-Skala von *Ablehnung stark* (1) bis *Zustimmung stark* (6) beantwortet, bei der auch die übrigen Skalenpunkte beschriftet waren und alle Skalenpunkte mit dem Ziel der Äquidistanz und einem einheitlichen Verständnis vorab näher erläutert wurden.

5.2.3 Durchführung

Die Erhebungen, an denen bis zu sechs Testpersonen gleichzeitig teilnehmen konnten, fanden überwiegend in ruhigen Laborräumen der Bergischen Universität Wuppertal statt. Wenige Erhebungen wurden auch in privaten Räumlichkeiten der Versuchsleiterinnen durchgeführt, wobei hier darauf geachtet wurde, zu den Laborräumen vergleichbare Bedingungen zu schaffen. Abgesehen von NEO-FFI und Modul 4 des WIT-2, die entsprechend den Angaben in den Manualen im Paper-Pencil-Format durchgeführt wurden (Borkenau & Ostendorf, 2008;

Kersting et al., 2008), fand die Erhebung computergestützt statt. Für die Programmierung und computergestützte Darbietung wurde Inquisit 4 (Millisecond Software, 2016) verwendet. Die auditive Präsentation der Instruktion zu den AcquA-PU Aufgaben und der Dialoge in den Aufgaben erfolgte über einen bereitgelegten Kopfhörer. Bei der computergestützten Präsentation ergab sich der einzige relevante Unterschied zwischen den Erhebungen an der Universität und im privaten Umfeld. Während in der Universität externe Monitore (23.8 Zoll) verwendet wurden, musste die Erhebung im privaten Umfeld an Laptops (15.6 Zoll) ohne externen Monitor durchgeführt werden.

Tabelle 5.2

Reihenfolge, Art der Administration und Zeitschätzung (in Minuten) der in Studie 2 eingesetzten Instrumente

Instrument	Administration	Zeitschätzung
1. Allgemeine Instruktion & Demografie	Computer	7
2. Modul 4 des WIT-2	Paper-Pencil	40
3. NEO-FFI	Paper-Pencil	10
4. AcquA-PU Teil 1 (Instruktion, 2 Aufgaben)[a]	Computer	30
Pause (ca. 5 Minuten; optional)		
5. Big Five Marker-Adjektive	Computer	10
6. AcquA-PU Teil 2 (4 Aufgaben)[a]	Computer	40
7. IPIP-AB5C-Items	Computer	10

Anmerkungen. Die geschätzte Studiendauer betrug insgesamt ca. 152 Minuten (inkl. Pause).
[a] Die Aufgaben wurden in einer unvollständig ausbalancierten Reihenfolge präsentiert. Hierfür wurden sechs mögliche Reihenfolgen erstellt, zu denen die Testpersonen zufällig zugewiesen wurden. Bei der Erstellung der Reihenfolgen wurde darauf geachtet, dass immer eine überarbeitete und eine neu konstruierte Aufgabe abwechselnd präsentiert wurden.

Zu Beginn jeder Erhebung erhielten die Testpersonen in schriftlicher Form umfangreiche allgemeine Information zur Studie. Zudem gaben die Testpersonen ihre schriftliche Einwilligung in die Studienteilnahme (bei minderjährigen Testpersonen wurde zusätzlich das Einverständnis einer erziehungsberechtigten Person eingeholt). Im Anschluss begann die Studie durch das Vorlesen der allgemeinen Instruktion durch die Versuchsleitung,

die die Testpersonen am Computerbildschirm verfolgen sollten. Der restliche Ablauf inklusive Zeitschätzungen kann Tabelle 5.2 entnommen werden. Auf Grund der einzuhaltenden Bearbeitungszeiten wurden die Aufgaben des WIT-2 in der Gruppe instruiert, die restlichen Aufgaben konnten von den Testpersonen in individueller Geschwindigkeit bearbeitet werden.

Nach Abschluss der Datenerhebung konnten die Testpersonen zwischen drei möglichen Aufwandsentschädigungen wählen: 1.) Erhalt von 12 € in bar; 2.) Rückmeldung zur eigenen Persönlichkeit auf Basis des NEO-FFI Selbstberichts; 3.) Versuchspersonenstunden entsprechend der tatsächlichen Studiendauer. Insgesamt variierte die Studiendauer der meisten Termine zwischen 2:00 h und 2:30 h.

5.2.4 Statistische Analysen

5.2.4.1 Allgemeines. Zur Überprüfung der faktoriellen Validität der AcquA-PU Aufgaben wie auch des Zusammenhangs zwischen PU und schlussfolgerndem Denken sowie PU und den Big Five wurden CFAs verwendet. Wie in der vorherigen Studie wurden in den Messmodellen für PU auf Grund der fehlenden lokalen stochastischen Unabhängigkeit der Items Aufgaben-Parcels als Indikatoren verwendet (vgl. Hellwig et al., 2020). Die Schätzung der CFA-Modelle erfolgte durch Mplus (Muthén & Muthén, 2017) sowie das R-Paket lavaan (Rosseel, 2012) und für die Schätzung der Reliabilitäten wurde das R-Paket MBESS (Kelley, 2007, 2020) verwendet. Als Konfidenzintervalle für Cronbachs α sowie McDonalds ω wurden – falls nicht anders angegeben – Bootstrap-Perzentil-Intervalle mit 10000 Replikationen bestimmt (vgl. Kelley, 2020; Kelley & Pornprasertmanit, 2016).

5.2.4.2 Itemselektion AcquA-PU Aufgaben. Für die Selektion geeigneter AcquA-PU Items wurde das in Abschnitt 4.2.5.2 beschriebene Vorgehen angewendet. Auf Grund der extremen Leichtigkeit einiger Items und Aufgaben, die sich in der vorherigen Studie insbesondere beim dichotomen Scoring deutlich zeigte (vgl. Tabelle A1 in Anhang A sowie Tabelle B1 in Anhang B), wurde allerdings folgende Anpassung vorgenommen: Beim dichotomen Scoring wurden in einem ersten Selektionsschritt alle Items mit Mittelwerten > .95 sowie < .05 eliminiert. Beim standardisierten Distanzscoring konnte diese Anpassung auf Grund der Itemscores, deren absolute Höhe nicht so eindeutig zu interpretieren ist wie beim dichotomen Scoring, nicht vorgenommen werden.

5.2.4.3 Zusammenhang PU und schlussfolgerndes Denken. Zur Untersuchung des Zusammenhangs zwischen PU und schlussfolgerndem Denken wurde ein zweifaktorielles CFA-Modell spezifiziert. Als Indikatoren für PU wurden die sechs Aufgaben-Parcels verwendet und als Indikatoren für das schlussfolgernde Denken die Summenwerte der drei Untertests des Moduls 4 aus dem WIT-2 (vgl. Kersting et al., 2008).

5.2.4.4 Zusammenhang PU und Persönlichkeit der Testperson. Der Zusammenhang zwischen PU und den Big Five wurde in einem separaten sechsfaktoriellen CFA-Modell geschätzt. Auf Grund der für die 60 Items des NEO-FFI verhältnismäßig kleinen Stichprobe wurden als Indikatoren der Big Five nicht die Items selbst, sondern hieraus gebildete Parcels verwendet. Je Faktor wurden drei Parcels erstellt, indem die 12 Items zufällig und gleichmäßig auf die Parcels aufgeteilt und anschließend gemittelt wurden (Little et al., 2002). Um die Stabilität der Ergebnisse und die Abhängigkeit von den Parcelzuteilungen zu überprüfen, wurden insgesamt drei zufällige Parcel-Sets gebildet, jeweils die Zusammenhänge zwischen PU und den Big Five geschätzt und anschließend die Ergebnisse miteinander verglichen.

Darüber hinaus sollte analog zu Studie 1 der Zusammenhang zwischen Skalen der verwendeten IPIP-Items von Goldberg (1999) und hypothesengeleitet gebildeten PU-Parcels untersucht werden, um einen möglichen Effekt der Ähnlichkeit zwischen Testperson und Zielperson zu explorieren. Um zwischen einer Skalenbildung auf Basis des AB5C-Algorithmus von Hofstee et al. (1992) und auf Basis von Faktorenanalysen zu entscheiden, wurde in einem ersten Schritt die Replizierbarkeit der in Studie 1 gebildeten AB5C-Skalen der Gewissenhaftigkeit überprüft. Hierfür erfolgte zunächst eine Kreuzvalidierung der CFA-Modelle der Marker-Adjektive von Ostendorf (1990) aus Studie 1. Im Anschluss wurden auf Basis der Daten aus Studie 2 die Faktorwerte der Big Five geschätzt, der AB5C-Algorithmus auf die IPIP-Items der Gewissenhaftigkeit angewendet (vgl. Abschnitt 4.2.5.4) und schließlich die gebildeten AB5C-Skalen aus Studie 1 und 2 miteinander verglichen. Im Falle einer erfolgreichen Replikation sollte eine entsprechende Skalenbildung mit den IPIP-Items der Verträglichkeit vorgenommen werden. Alternativ wurde eine Skalenbildung auf Basis exploratorischer und konfirmatorischer Faktorenanalysen geplant.

Abschließend wurden die aus den IPIP-Items gebildeten Skalen mit hypothesengeleitet gebildeten PU-Parcels korreliert. Während in der vorherigen Studie diejenigen PU-Items zusammengefasst wurden, die sich auf Zielpersonen mit derselben relevanten Kontingenz beziehen (d.h. deren Scoring-Rationale auf demselben IPIP-AB5C-Item basiert), wurden in der vorliegenden Studie alle Items zusammengefasst, die sich unabhängig von der spezifischen

Kontingenz auf Zielpersonen mit derselben Ausprägung einer Persönlichkeitseigenschaft beziehen. Das heißt, es wurden beispielsweise alle Items zusammengefasst, mit denen eine Zielperson eingeschätzt werden sollte, in deren relevanter Kontingenz eine hohe Ausprägung der Gewissenhaftigkeit modelliert wurde. Dieses abgeänderte Vorgehen wurde gewählt, da die Parcels ansonsten sehr wenige Items enthalten hätten. In der vorherigen Studie standen je relevanter Kontingenz zwei Zielpersonen zur Verfügung, während in der aktuellen Studie je Kontingenz nur noch eine Zielperson eingeschätzt wurde. Da die Reliabilitäten der Parcels in Studie 1 bereits überwiegend zu niedrig ausfielen, wurde eine analoge Parcelbildung als nicht sinnvoll erachtet und das alternative Vorgehen gewählt. Insgesamt wurden vier hypothesengeleitete PU-Parcels gebildet: Die zusammengefassten Items des Parcels beziehen sich auf eine Zielperson mit 1.) hoher Ausprägung auf dem Faktor Gewissenhaftigkeit, 2.) niedriger Ausprägung auf dem Faktor Gewissenhaftigkeit, 3.) hoher Ausprägung auf dem Faktor Verträglichkeit, 4.) niedriger Ausprägung auf dem Faktor Verträglichkeit.

5.2.4.5 Power. Zur Überprüfung der Angemessenheit der Stichprobengröße wurde die Power für die zentralen Analysen der Studie bestimmt. Die Schätzung für das eindimensionale Messmodell von PU erfolgte unter Annahme von sechs Indikatoren mit Hilfe des von Preacher und Coffman (2006) zur Verfügung gestellten R-Code Generators. Für den χ^2-Test des absoluten Fits mit H_0: RMSEA = 0, α = .05, angenommener Effekt: RMSEA = .10, df = 9 und einem N von 204 ergab sich eine Power von .87. Unter Verwendung des Monte Carlo Ansatzes in Mplus (Muthén & Muthén, 2002, 2017) wurde zudem die Power für den ebenfalls im Fokus stehenden Zusammenhang zwischen PU und schlussfolgerndem Denken auf Basis des zweifaktoriellen CFA-Modells geschätzt (10000 Replikationen). Für die Schätzung wurden standardisierte Ladungen von .60 bei PU (konservative Annahme auf Grund der Überarbeitungen und Neukonstruktionen; durchschnittliche Ladung der PU-Aufgaben aus Studie 1 unter Verwendung des standardisierten Distanzscorings lag bei .73) und standardisierte Ladungen von .74 für die Untertests aus Modul 4 des WIT-2 (durchschnittliche Ladung der Untertests im Manual des WIT-2; Kersting et al., 2008) verwendet. Auf Basis der im zweiten Studienziel formulierten Annahme zum Zusammenhang zwischen PU und logischem Schlussfolgern wurde dieser in der Poweranalyse auf .50 festgelegt. Hieraus resultierte bei einem N von 204 eine Power von 1. Auch ein mittlerer Effekt von .30 würde mit einer Power von .92 entdeckt werden, sodass die Poweranalysen dafür sprechen, dass eine ausreichend große Stichprobe erhoben wurde.

5.3 Ergebnisse

Die im Folgenden präsentierten Ergebnisse wurden unter Verwendung des standardisierten Distanzscorings bei den AcquA-PU Aufgaben ermittelt. Die Ergebnisse unter Verwendung des dichotomen Scorings sind in Anhang C zu finden und werden nur zusammenfassend und insbesondere bei relevanten Abweichungen zum standardisierten Distanzscoring erwähnt.

5.3.1 AcquA-PU Aufgaben: Itemanalysen und Itemselektion

Die Betrachtung von Schiefe, Kurtosis sowie Histogrammen der Itemscores der AcquA-PU Aufgaben ließ die Annahme zu, dass bei einer Vielzahl der Items keine univariate Normalverteilung und folglich bei keiner Aufgabe eine multivariate Normalverteilung vorlag (Range Schiefe: -2.15 bis 5.59; 22 der 84 Items zeigten eine Schiefe > 2; Range Kurtosis: -1.55 bis 53.47). Im Rahmen der Itemselektion wurde für die Schätzung der CFA-Modelle daher der MLM-Schätzer mit Satorra-Bentler χ^2 verwendet (Muthén & Muthén, 2017).

Die verwendete Itemselektionsstrategie führte bei Aufgabe PU3 zur Elimination von 17 der ursprünglich 24 Items. Auffällig war hierbei, dass von den selektierten Items drei Items hohe standardisierte Ladungen zwischen .79 und .88 aufwiesen und vier Items vergleichsweise niedrige Ladungen zwischen .15 und .38. Auf Grund der hohen Anzahl an eliminierten Items wurde eine zweite Selektion vorgenommen, bei der zunächst die drei Items mit den hohen Ladungen eliminiert wurden. Diese zweite Itemselektion führte zur Elimination von 14 der 24 Items. Die im Anschluss an beide Selektionen gebildeten Parcels der Aufgabe PU3 zeigten im finalen Messmodell für PU gute standardisierte Ladungen (Selektion 1: .70; Selektion 2: .62) und die jeweiligen Messmodelle einen gleich guten Fit. Die Reliabilität auf Basis des finalen Messmodells fiel bei Verwendung des Parcels der zweiten Selektion geringfügig besser aus (Selektion 1: $\alpha = .54$, $\omega = .56$; Selektion 2: $\alpha = .57$, $\omega = .59$). Auffällig war jedoch, dass bei der zweiten Selektion Aufgabe PU6 im Messmodell eine bessere Ladung aufwies (Selektion 1: .28; Selektion 2: .49). Unter Berücksichtigung aller Ergebnisse wurde daher im Folgenden das Ergebnis der zweiten Selektion verwendet.

Die Itemselektionen bei den Aufgaben PU4 und PU5 waren ebenfalls auffällig. In beiden Fällen wurde unter anderem eine komplette Anwendungsphase eliminiert, da die entsprechenden Items Ladungen nahe 0 aufwiesen.

Insgesamt wurde die Itemanzahl von 84 auf 45 reduziert. Deskriptive Statistiken der selektierten Items sind in Tabelle D1 in Anhang D zu finden. Die part-whole korrigierten Trennschärfen dieser Items, berechnet über die Items aller Aufgaben hinweg, liegen bei .03 bis

.43 (*M* = .25, *SD* = .10) und die Itemmittelwerte bei 0.22 bis 1.49 (*M* = 0.62, *SD* = 0.32). Hierbei ist zu beachten, dass auf Grund der standardisierten Distanzen höhere Itemmittelwerte eine höhere Schwierigkeit anzeigen.

Deskriptive Statistiken der im Anschluss an die Itemselektion gebildeten aufgabenweisen Parcels und eines PU-Gesamtscores finden sich in Tabelle 5.3. Bei den Parcels und dem Gesamtscore handelt es sich um Mittelwerte der jeweiligen Items.

Tabelle 5.3

Deskriptive Statistiken der Personality Understanding (PU)-Parcels sowie des gesamten AcquA-PU Tests (PU_{ges}) nach der Itemselektion (standardisiertes Distanzscoring)

Parcel	Itemanzahl[a]	*M*	*SD*	Schiefe[b]	Kurtosis[c]
PU1	10 (12)	0.44	0.17	0.83	1.65
PU2	7 (12)	0.54	0.39	1.69	2.33
PU3	10 (24)	0.49	0.18	1.02	4.11
PU4	6 (12)	0.64	0.44	1.61	1.92
PU5	5 (12)	1.28	0.54	-0.29	-1.18
PU6	7 (12)	0.63	0.21	0.75	1.55
PU_{ges}	45 (84)	0.62	0.15	0.02	-0.08

Anmerkungen. N = 204. Die verwendete Statistik zur Schätzung der Kurtosis nimmt beim Vorliegen einer Normalverteilung den Wert 0 an.

[a] In Klammern ist die Anzahl vor der Itemselektion angegeben. [b] SE = 0.17. [c] SE = 0.34.

Im Vergleich zum standardisierten Distanzscoring musste unter Verwendung des dichotomen Scorings eine noch größere Anzahl an Items eliminiert werden, sodass nur drei bis sechs Items je Aufgabe für die Parcelbildung verwendet werden konnten (vgl. Abschnitt C1 und Tabelle C1 in Anhang C). Es zeigten sich dabei ähnliche Probleme wie beim standardisierten Distanzscoring, da beim dichotomen Scoring ebenfalls die Anwendungsphasen der Aufgaben PU4 und PU5 eliminiert wurden. Tabelle D1 in Anhang D macht zudem deutlich, dass die selektieren Items des dichotomen Scorings überwiegend auch beim standardisierten Distanzscoring selektiert wurden. Ausnahmen bilden lediglich zwei Items der Aufgabe PU3, ein Item der Aufgabe PU5 sowie zwei Items der Aufgabe PU6.

5.3.2 AcquA-PU Aufgaben: Messmodell und Reliabilität

Unter Verwendung der zuvor gebildeten Aufgaben-Parcels wurde das eindimensionale Messmodell für PU geschätzt. Da Schiefe, Kurtosis sowie Histogramme der Parcels Abweichungen von einer univariaten Normalverteilung einiger Indikatoren aufzeigten, wurde erneut der MLM-Schätzer verwendet. Das angenommene Modell zeigte mit $\chi^2 = 9.37$, $df = 9$, $p = .40$, CFI = .996, RMSEA = .01, 90% CI = [.00, .08] einen sehr guten Fit ($N = 204$). Die standardisierten Faktorladungen, die für die Aufgaben PU4 und PU5 sehr gering und nicht signifikant ausfielen, sind in Tabelle 5.4 zu finden.

Tabelle 5.4

Ergebnis der konfirmatorischen Faktorenanalyse aller sechs Personality Understanding (PU)-Parcels (standardisiertes Distanzscoring)

Parcel	standardisierte Faktorladung
PU1	.61***
PU2	.50***
PU3	.65***
PU4	.14
PU5	.09
PU6	.49***

Anmerkungen. $N = 204$.

***$p < .001$.

Die Schätzung der Reliabilität erfolgte auf Grund der vorliegenden Eindimensionalität unter Verwendung von Cronbachs α sowie McDonalds ω. Auf Basis der sechs Parcels resultierte ein Wert für Cronbachs α von .36, 95% CI = [.21, .48] und für McDonalds ω von .37, 95% CI = [.24, .48]. Auf Grund der geringen Reliabilitäten wurde untersucht, ob eine Elimination der Aufgaben PU4 und PU5 die Reliabilitätsschätzungen verbessert. Der schrittweise Ausschluss beider Aufgaben führte zu einer substantiellen Erhöhung der Reliabilität, auch wenn die neuen Schätzungen immer noch in keinem akzeptablen Bereich lagen: α = .57, 95% CI = [.47, .65]; ω = .59, 95% CI = [.48, .69]. Auch das Messmodell wurde auf Basis der vier verbliebenen Aufgaben neu geschätzt und zeigte einen perfekten Fit (MLM-Schätzer): $\chi^2 = 1.16$, $df = 2$, $p = .56$, CFI = 1, RMSEA = 0, 90% CI = [.00, .12] ($N = 204$, standardisierte Faktorladungen siehe Tabelle 5.5).

Tabelle 5.5

Ergebnis der konfirmatorischen Faktorenanalyse der Personality Understanding (PU)-Parcels PU1, PU2, PU3 und PU6 (standardisiertes Distanzscoring)

Parcel	standardisierte Faktorladung
PU1	.63***
PU2	.50***
PU3	.62***
PU6	.49***

Anmerkungen. $N = 204$.

***$p < .001$.

Auf Grund der besseren Ergebnisse im Zusammenhang mit der Reliabilität wurde das Modell unter Ausschluss der Aufgaben PU4 und PU5 als finales Messmodell verwendet.

Bei Anwendung des dichotomen Scorings wurden ebenfalls die Aufgaben PU4 und PU5 eliminiert. Im Hinblick auf die Ladungen der Aufgaben-Parcels im finalen Messmodell sowie bei den Reliabilitätsschätzungen ergaben sich zwischen beiden Scoring-Methoden zum Teil deutliche Unterschiede (vgl. Abschnitt C.2 in Anhang C). Die standardisierten Ladungen fielen beim dichotomen Scoring bei allen vier Aufgaben mit .22 (PU6) bis .55 (PU1) geringer aus als bei Verwendung des standardisierten Distanzscorings. Folglich resultierten mit $\alpha = .39$ und $\omega = .40$ auch geringere Reliabilitätsschätzungen.

5.3.3 WIT-2: Deskriptive Statistiken und Reliabilität

Entsprechend den Vorgaben im Manual des WIT-2 (Kersting et al., 2008) wurden bei der Auswertung der drei Untertests Abwicklungen, Analogien und Zahlenreihen Summenwerte der richtig gelösten Items sowie ein Gesamtsummenwert für die Skala schlussfolgerndes Denken gebildet. Deskriptive Statistiken der drei Untertests sowie deskriptive Statistiken und Reliabilitätsschätzungen der Gesamtskala finden sich in Tabelle 5.6. Zum Vergleich mit den Ergebnissen im Manual wurde Cronbachs α auf Itemebene bestimmt. Zusätzlich wurde McDonalds ω unter Verwendung der drei Summenscores der Untertests geschätzt, da diese in den folgenden Strukturanalysen als Indikatoren verwendet wurden.

Cronbachs α fiel mit .91 nur geringfügig niedriger aus als im Manual des WIT-2, wo ein stratifiziertes α von .94 berichtet wird (Kersting et al., 2008). Zudem fiel die Schätzung der Reliabilität auf Parcelebene geringer aus als auf Itemebene ($\omega = .72$ vs. $\alpha = .91$).

Tabelle 5.6

*Itemanzahl (*k*) und deskriptive Statistiken der eingesetzten Untertests und der Skala schlussfolgerndes Denken des WIT-2 sowie Reliabilitätsschätzungen (Cronbachs α, McDonalds ω) für das schlussfolgernde Denken*

Skala/Untertest	*k*	*M*	*SD*	Schiefe[a]	Kurtosis[b]	α [95% CI][c]	ω [95% CI][d]
Schlussfolgerndes Denken	60	32.02	10.23	0.06	-0.53	.91 [.89, .92]	.72 [.65, .78]
Abwicklungen	20	11.26	4.47	-0.14	-0.71		
Analogien	20	11.25	4.03	0.08	-0.65		
Zahlenreihen	20	9.50	4.30	0.04	-0.37		

Anmerkungen. N = 204. CI = Konfidenzintervall. Die verwendete Statistik zur Schätzung der Kurtosis nimmt beim Vorliegen einer Normalverteilung den Wert 0 an.

[a] SE = 0.17. [b] SE = 0.34. [c] Schätzung auf Itemebene. [d] Schätzung auf Parcelebene (drei Parcels).

5.3.4 Zusammenhang PU und schlussfolgerndes Denken

Für das zweifaktorielle CFA-Modell zur Untersuchung des Zusammenhangs zwischen PU und schlussfolgerndem Denken wurde auf Grund der Verteilungen der PU-Parcels erneut der MLM-Schätzer verwendet. Als Indikatoren für PU wurden anders als ursprünglich geplant nur die vier Aufgaben PU1, PU2, PU3 und PU6 des finalen Messmodells verwendet.

Das Modell zeigte mit χ^2 = 5.44, df = 13, p = .96, CFI = 1, RMSEA = 0, 90% CI = [.00, .00] einen perfekten Fit auf die Daten (N = 204). Die restlichen Ergebnisse der CFA sind in Tabelle 5.7 dargestellt und zeigen einen kleinen und nicht signifikanten Zusammenhang zwischen den beiden Faktoren. Bei der Interpretation der Faktorkorrelation muss beachtet werden, dass bei PU auf Grund des standardisierten Distanzscorings ein niedrigerer Wert eine bessere Leistung anzeigt. Der Zusammenhang deutet rein deskriptiv betrachtet somit darauf hin, dass eine höhere Ausprägung auf dem Faktor PU tendenziell mit einer höheren Fähigkeit zum schlussfolgernden Denken einhergeht.

Die Ergebnisse unter Verwendung des dichotomen Scorings finden sich in Tabelle C4 in Anhang C. Hier fiel der latente Zusammenhang zwischen PU und schlussfolgerndem Denken mit .09 noch geringer aus als bei Verwendung des standardisierten Distanzscorings. Auffällig ist hier zudem das große 95%-Konfidenzintervall der latenten Korrelation, das von -.17 bis .35 reicht.

Tabelle 5.7

Ergebnis der konfirmatorischen Faktorenanalyse der Personality Understanding (PU)-Parcels (standardisierte Distanzen) und Parcels des schlussfolgernden Denkens (R)

	standardisierte Faktorladung	
Parcel	PU	R
PU1	.62***	
PU2	.51***	
PU3	.64***	
PU6	.48***	
Al		.62***
Zn		.65***
Aw		.76***
Faktorkorrelation [95% CI]		
PU - R	-.22 [-.45, .00][a]	

Anmerkungen. N = 204. Al = Analogien; Zn = Zahlenreihen; Aw = Abwicklungen; CI = Konfidenzintervall.

***p < .001. [a] Konfidenzintervall überdeckt den Wert 0.

5.3.5 NEO-FFI: Deskriptive Statistiken und Reliabilität

Deskriptive Statistiken und Reliabilitätsschätzungen der NEO-FFI Skalen Neurotizismus, Extraversion, Offenheit für Erfahrungen, Verträglichkeit und Gewissenhaftigkeit sind in Tabelle 5.8 dargestellt. Wie bei den Untertests des WIT-2 wurde die Schätzung der Reliabilität einerseits zum Vergleich mit den Angaben im Manual mittels Cronbachs α auf Itemebene vorgenommen und andererseits mittels McDonalds ω auf Ebene derjenigen Parcels, die in den folgenden Strukturanalysen als Indikatoren verwendet wurden.

Die Ergebnisse für Cronbachs α sind überwiegend vergleichbar mit den Angaben im Manual des NEO-FFI (betrachtet wurden hier die Ergebnisse auf Basis der Gesamtstichprobe). Der Wert für Verträglichkeit fiel in der vorliegenden Studie mit α = .80 sogar etwas höher aus (α = .72 im Manual; Borkenau & Ostendorf, 2008).

Tabelle 5.8

*Itemanzahl (*k*), deskriptive Statistiken sowie Reliabilitätsschätzungen (Cronbachs α, McDonalds ω) der Skalen des NEO-FFI*

Skala	*k*	*M*	*SD*	Schiefe[a]	Kurtosis[b]	α [95% CI][c]	ω [95% CI][d]
N	12	23.30	8.79	0.26	-0.45	.87 [.84, .89]	.88 [.86, .91][e]
E	12	27.85	7.09	-0.47	0.22	.80 [.75, .84]	.83 [.78, .88]
O	12	32.73	6.49	-0.34	-0.19	.74 [.69, .79]	.77 [.71, .82]
A	12	32.63	6.85	-0.66	0.67	.80 [.74, .84]	.81 [.75, .86]
C	12	32.06	7.43	-0.33	-0.30	.85 [.81, .87]	.84 [.80, .88]

Anmerkungen. $N = 204$. N = Neurotizismus; E = Extraversion; O = Offenheit für Erfahrungen; A = Verträglichkeit; C = Gewissenhaftigkeit; CI = Konfidenzintervall. Die verwendete Statistik zur Schätzung der Kurtosis nimmt beim Vorliegen einer Normalverteilung den Wert 0 an.
[a] $SE = 0.17$. [b] $SE = 0.34$. [c] Schätzung auf Itemebene. [d] Schätzung auf Parcelebene (Parcelset 1). Die Ergebnisse aller drei Parcelsets finden sich in Tabelle D2 in Anhang D. [e] Bei Bestimmung des Konfidenzintervalls kam es wiederholt zu Schätzproblemen (insb. negative Varianzen).

5.3.6 Zusammenhang PU und Big Five

Die Ergebnisse des sechsfaktoriellen CFA-Modells zur Schätzung des Zusammenhangs zwischen den PU und den Big Five-Faktoren (Parcel-Set 1) sind in den Tabellen 5.9 und 5.10 dargestellt (für die Ergebnisse unter Verwendung der Parcel-Sets 2 und 3 siehe Tabelle D3 und D4 in Anhang D). Als Indikatoren für PU wurden erneut nur die vier Aufgaben PU1, PU2, PU3 und PU6 verwendet. Als Schätzer wurde der MLM-Schätzer gewählt.

Bei der Interpretation der Faktorkorrelationen in Tabelle 5.10 muss beachtet werden, dass bei PU auf Grund der standardisierten Distanzen ein niedrigerer Wert eine bessere Leistung anzeigt. Insgesamt resultierten maximal kleine und nicht signifikante Korrelationen zwischen PU und den Big Five. Die Höhe der Zusammenhänge variiert zudem geringfügig in Abhängigkeit des verwendeten Parcel-Sets. Die Ergebnisse unter Verwendung des dichotomen Scorings sind in Tabelle C5 bis C7 (Anhang C) zu finden und unterscheiden sich nur punktuell und geringfügig von den Ergebnissen des standardisierten Distanzscorings. Auch hier ergaben sich maximal geringe und überwiegend nicht signifikante Zusammenhänge zwischen PU und den Big Five. Darüber hinaus zeigten alle Modelle beider Scoring-Methoden nur einen akzeptablen, aber entsprechend der Kriterien von Browne und Cudeck (1992) sowie Hu und Bentler (1999) keinen guten Modell-Fit.

Tabelle 5.9

Ergebnis der konfirmatorischen Faktorenanalyse der Personality Understanding (PU)-Parcels (standardisiertes Distanzscoring) und NEO-FFI-Parcels (Parcel-Set 1)

	standardisierte Faktorladung					
Parcel	PU	N	E	O	A	C
PU1	.63***					
PU2	.51***					
PU3	.61***					
PU6	.49***					
nffi_n1a		.74***				
nffi_n1b		.94***				
nffi_n1c		.81***				
nffi_e1a			.92***			
nffi_e1b			.70***			
nffi_e1c			.70***			
nffi_o1a				.82***		
nffi_o1b				.59***		
nffi_o1c				.73***		
nffi_a1a					.80***	
nffi_a1b					.65***	
nffi_a1c					.84***	
nffi_c1a						.85***
nffi_c1b						.84***
nffi_c1c						.69***

Anmerkungen. $N = 204$. N = Neurotizismus; E = Extraversion; O = Offenheit für Erfahrungen; A = Verträglichkeit; C = Gewissenhaftigkeit. Modell-Fit: $\chi^2 = 251.90$, $df = 137$, $p < .001$, CFI = .92, RMSEA = .06, 90% CI = [.05, .08] (MLM-Schätzer). Faktorkorrelationen siehe Tabelle 5.10.

***$p < .001$.

Tabelle 5.10

Faktorkorrelationen zwischen Personality Understanding (PU) und den Big Five-Faktoren

	PU	N	E	O	A
N	-.01 [-.18, .16]				
E	-.06 [-.22, .10]	-.39 [-.52, -.27]			
O	-.15 [-.31, .01]	-.03 [-.18, .12]	.10 [-.06, .26]		
A	-.13 [-.30, .05]	.05 [-.10, .19]	.52 [.42, .62]	.02 [-.12, .16]	
C	.14 [-.03, .31]	-.26 [-.41, -.11]	.28 [.13, .43]	-.15 [-.30, -.01]	.17 [-.00, .35]

Anmerkungen. $N = 204$. N = Neurotizismus; E = Extraversion; O = Offenheit für Erfahrungen; A = Verträglichkeit; C = Gewissenhaftigkeit. Für die verwendeten Indikatoren sowie den Modell-Fit siehe Tabelle 5.9. In eckigen Klammern sind die 95%-Konfidenzintervalle der Korrelationen dargestellt.

Latente Korrelationen ≥ |.10| finden sich über alle Modelle beider Scoring-Methoden hinweg insbesondere für die Faktoren Offenheit für Erfahrungen, Gewissenhaftigkeit und Verträglichkeit. Bei Offenheit und Verträglichkeit geht eine höhere Ausprägung auf der Persönlichkeitseigenschaft mit einer besseren Leistung bei PU einher – bei Gewissenhaftigkeit liegt das umgekehrte Muster vor. Für einige Parcel-Bildungen resultierten auch kleine Zusammenhänge zwischen PU und Extraversion.

5.3.7 PU-Parcels und Skalen der IPIP-Items

5.3.7.1 Kreuzvalidierung Marker-Adjektive. Zunächst wurden die einfaktoriellen Messmodelle der Big Five-Faktoren sowie das fünffaktorielle Gesamtmodell aus Studie 1 (drei Parcel-Sets; vgl. Abschnitt 4.3.6.1) auf Basis der Daten aus Studie 2 neu geschätzt. Auf Grund der univariaten Verteilungen der Adjektive und Adjektiv-Parcels wurde der MLM-Schätzer verwendet. Die Ergebnisse sind in Tabelle 5.11 dargestellt und zeigen, außer für Verträglichkeit sowie knapp für Offenheit für Erfahrungen, keinen guten Fit der einzelnen Messmodelle entsprechend den Cut-Off-Kriterien von Browne und Cudeck (1992) sowie Hu und Bentler (1999). Zudem fiel der Fit für alle fünf Messmodelle schlechter aus als in Studie 1 (vgl. Tabelle A4, Anhang A). Für das fünffaktorielle Gesamtmodell ergaben sich hingegen Ergebnisse, die vergleichbar sind mit denen der vorherigen Studie (vgl. Abschnitt 4.3.6.1). Es erschien daher angemessen, die Gesamtmodelle zur Schätzung der Faktorwerte zu verwenden.

Tabelle 5.11

Ergebnisse der konfirmatorischen Faktorenanalysen der Adjektive von Ostendorf (1990) zur Kreuzvalidierung der einfaktoriellen Messmodelle der Big Five-Faktoren sowie des fünffaktoriellen Gesamtmodells aus Studie 1 auf Basis der Daten aus Studie 2

	χ^2				
Modell	Wert	*df*	*p*	CFI	RMSEA [90% CI]
Messmodell Offenheit für Erfahrungen	51.96	27	.003	.94	.07 [.04, .10]
Messmodell Gewissenhaftigkeit	149.36	65	< .001	.90	.08 [.06, .10]
Messmodell Extraversion	112.81	35	< .001	.88	.10 [.08, .13]
Messmodell Verträglichkeit	85.83	54	.004	.95	.05 [.03, .07]
Messmodell Neurotizismus	67.91	27	< .001	.86	.09 [.06, .11]
Gesamtmodell Parcel-Set 1	134.29	80	< .001	.96	.06 [.04, .07]
Gesamtmodell Parcel-Set 2	187.60	80	< .001	.92	.08 [.07, .10]
Gesamtmodell Parcel-Set 3	160.36	80	< .001	.94	.07 [.05, .09]

Anmerkungen. N = 204. *df* = Freiheitsgrade; CFI = Comparative Fit Index; RMSEA = Root-Mean-Sqaure Error of Approximation; CI = Konfidenzintervall.

5.3.7.2 AB5C-Skalenbildung: Vergleich Studie 1 und Studie 2. Schließlich wurden auf Basis der Daten der vorliegenden Studie 2 die Faktorwerte der Big Five geschätzt und der AB5C-Algorithmus auf die IPIP-Items der Gewissenhaftigkeit angewendet. Die resultierenden AB5C-Skalen wurden im Anschluss mit denen der vorherigen Studie verglichen. Hierbei zeigte sich, dass 12 der insgesamt 34 Gewissenhaftigkeits-Items in beiden Studien derselben AB5C-Skala zugeordnet wurden. Die restlichen 22 Items wurden in Studie 2 allerdings einer anderen AB5C-Skala zugeordnet als in Studie 1, sodass der überwiegende Teil der AB5C-Skalen nicht repliziert werden konnte. Zudem war in der vorliegenden Studie analog zu Studie 1 in vielen Fällen auf Grund von ähnlich hohen Nebenladungen keine eindeutige Skalenzuordnung möglich, sodass die vorgenommene Zuordnung beliebig erschien. Insgesamt und über beide Studien hinweg betrachtet, sprechen die Ergebnisse somit gegen eine Skalenbildung auf Basis des AB5C-Algorithmus. Daher wurde für die folgenden Analysen eine alternative Skalenbildung auf Basis von Faktorenanalysen vorgenommen, die im Folgenden näher beschrieben wird.

5.3.7.3 Alternative Skalenbildung IPIP-Items. Die alternative Skalenbildung wurde für die IPIP-Items der Gewissenhaftigkeit und die der Verträglichkeit leicht unterschiedlich vorgenommen. Da die Items der Gewissenhaftigkeit auch in der vorherigen Studie erhoben wurden, lagen hier Daten für eine deutlich höhere Anzahl an Personen vor. Diese wurden zunächst zusammengefügt (N = 403 vollständige Datensätze) und anschließend zufällig auf zwei Datensätze (n_1 = 252, n_2 = 151) aufgeteilt. Auf Basis des größeren Datensatzes wurde mittels exploratorischer Faktorenanalysen (EFAs) die interne Struktur der Items untersucht, geeignete Items selektiert und eine Skalenbildung vorgenommen. Diese wurde anschließend auf Basis des zweiten Datensatzes mit Hilfe einer CFA kreuzvalidiert. Für die Items der Verträglichkeit wurde die Skalenbildung auf Grund der geringeren Datenmenge (nur Studie 2; N = 204) ausschließlich mittels EFAs vorgenommen.

Gewissenhaftigkeit. Die EFAs wurden als Maximum-Likelihood-EFAs (ML-EFAs) durchgeführt. Auf Grund der univariaten Verteilungen der Gewissenhaftigkeits-Items wurde der MLR-Schätzer verwendet, ein ML-Schätzer mit gegenüber Verletzungen der Annahme einer multivariaten Normalverteilung robusten Standardfehlern sowie χ^2-Teststatistik (Muthén & Muthén, 2017). Um die Anzahl der zu extrahierenden Faktoren zu bestimmen, wurden entsprechend den Empfehlungen von Fabrigar et al. (1999) mehrere Kriterien betrachtet. Zum einen wurde eine Parallelanalyse mit Hilfe der von O'Conner (2000) erstellten SPSS-Syntax durchgeführt. Verwendet wurden hierbei die mittels Hauptachsenanalyse gewonnenen Eigenwerte. Zusätzlich wurde der entsprechende Screeplot der geschätzten empirischen Eigenwerte erstellt. Da primär ML-EFAs durchgeführt wurden, konnte darüber hinaus auch der Fit für EFA-Modelle mit verschieden vielen Faktoren betrachtet werden.

Die Parallelanalyse unter Verwendung aller 34 Items deutete darauf hin, dass sechs Faktoren extrahiert werden sollten. Dies ergab ein Vergleich der empirischen Eigenwerte mit dem .95-Quantil der Eigenwerte von 5000 simulierten zufälligen Datensätzen. Auch der Fit des sechsfaktoriellen EFA-Modells fiel akzeptabel bis gut aus (χ^2 = 597.64, df = 372, p < .001, CFI = .94, RMSEA = .05, 90% CI = [.04, .06]) sowie signifikant besser als der Fit eines fünffaktoriellen Modells ($\tilde{T}_d$= 52.93, df = 29, p = .004; skalierte Teststatistik aus Mplus nach Satorra & Bentler, 2001; vgl. Muthén & Muthén, 2017), wobei die Unterschiede in den deskriptiven Fit-Indizes gering waren (ΔCFI = -.01, ΔRMSEA = .00). Der Screeplot ergab einen uneindeutigen Eigenwertverlauf. Eine inhaltliche Betrachtung des oblique Geomin-

rotierten[11] Ladungsmusters zeigte zudem, dass die sechsfaktorielle Lösung zu einem sechsten Faktor führt, auf dem nur zwei Items Primärladungen ≥ |.30| aufweisen, die darüber hinaus inhaltlich weitestgehend redundant sind: „Ich plane gerne im Voraus“ (Item 32, Ladung: .97), „Ich plane nicht im Voraus“ (Item 24, Ladung: -.63). Zur Vermeidung eines solchen Methodenfaktors wurde das Item mit der höchsten Ladung auf diesem Faktor eliminiert und die Analysen mit dem reduzierten Itempool, das heißt ohne Item 32, wiederholt.

Eine Parallelanalyse unter Verwendung der 33 Items wies auf fünf zugrunde liegende Faktoren hin. Der Fit der fünffaktoriellen Lösung fiel akzeptabel bis gut aus (χ^2 = 595.60, *df* = 373, $p < .001$, CFI = .94, RMSEA = .05, 90% CI = [.04, .06]). Die sechsfaktorielle Lösung zeigte nun auch keinen signifikant besseren Fit als die fünffaktorielle Lösung ($\tilde{T}_d$= 33.88, *df* = 28, $p = .20$). Der Screeplot blieb allerdings weiterhin uneindeutig. Eine inhaltliche Betrachtung des Ladungsmusters ergab erneut Zweifel an der Relevanz des letzten Faktors. Hier zeigten nur drei Items Primärladungen ≥ |.30|: „Ich erledige Dinge nach Vorschrift“ (Item 21, Ladung: .99), „Ich komme oft zu spät zur Arbeit“ (Item 34, Ladung: -.31), „Ich achte darauf, dass Regeln eingehalten werden“ (Item 25, Ladung: .41). Während Item 21 auf den anderen vier Faktoren Ladungen von maximal |.02| aufwies, zeigten die anderen beiden Items Sekundärladungen > |.25|. Um zu überprüfen, ob Faktor 5 primär auf Grund von Item 21 resultierte, wurde dieses in einem nächsten Schritt eliminiert.

Auf Basis der 32 Items ergab die Parallelanalyse vier zu extrahierende Faktoren (vgl. Tabelle D5, Anhang D; Vergleich empirische Eigenwerte mit dem .95-Quantil der Eigenwerte der Zufallsdaten) und auch der Screeplot (vgl. Abbildung D1, Anhang D) war mit vier Faktoren zu vereinbaren. Der Fit des vierfaktoriellen Modells fiel mit χ^2 = 564.80, *df* = 374, $p < .001$, CFI = .95, RMSEA = .05, 90% CI = [.04, .05] gut und nicht signifikant schlechter als der eines fünffaktoriellen Modells aus ($\tilde{T}_d$= 38.79, *df* = 28, $p = .08$). Eine inhaltliche Betrachtung des Ladungsmusters ergab zudem inhaltlich sinnvoll interpretierbare Faktoren.

Nach Festlegung auf vier zu extrahierende Faktoren wurde eine Itemselektion vorgenommen. Ziel war der Ausschluss von Items, die auf keinem Faktor eine substantielle Ladung aufweisen. Die Grenze wurde auf eine Ladung von ≥ |.30| gelegt. Des Weiteren sollte der Fit des EFA-Modells verbessert werden, sodass – analog zu den anderen vorgenommenen Itemselektionen – sowohl ein CFI von ≥ .95 als auch ein RMSEA von ≤ .05 erreicht wird (vgl.

[11] Neben der obliquen Geomin-Rotation wurden auch die oblique Varimax-Rotation sowie die Quartimin-Rotation exploriert. Alle drei Rotationen kamen zu ähnlichen Ergebnissen und wichen nur bei einzelnen Items voneinander ab. Insgesamt zeigte die Geomin-rotierte Lösung die besten Ergebnisse bezüglich Interpretation und Replikation der Skalenzuordnung, sodass diese verwendet wurde.

Browne & Cudeck, 1992; Hu & Bentler, 1999). Die Itemselektion wurde in einem iterativen Prozess durchgeführt, was heißt, dass nach jeder Elimination eines Items die EFA neu durchgeführt und auch die Dimensionalität neu überprüft wurde. Zunächst wurden Items eliminiert, die auf keinem Faktor eine Ladung ≥ |.30| zeigten, beginnend mit der absolut geringsten Primärladung. Dies führte zur Elimination eines Items. Anschließend sollten Items mit Nebenladungen ≥ |.30| eliminiert werden, sofern deren Elimination zu einer Verbesserung im Fit führt. Dies führte ebenfalls zur Elimination eines Items. Eine weitere Elimination war nicht erforderlich, da die Kriterien für den Modellfit erreicht wurden und zudem alle selektierten Items genau eine Ladung ≥ |.30| auf nur einem Faktor aufwiesen. Schließlich wurden die selektierten 30 Items entsprechend ihrer höchsten Ladung einem der vier Faktoren zugeordnet (vgl. Tabelle D6 in Anhang D). Die Inhalte der zugeordneten Items deuten darauf hin, dass Faktor 1 die Einhaltung von Regeln und Verpflichtungen repräsentiert, Faktor 2 planvolles Handeln, Faktor 3 (fehlende) Prokrastination und Faktor 4 Vorliebe für Ordnung.

Im nächsten Schritt folgte die Kreuzvalidierung der Zuordnung auf Basis des zweiten Datensatzes. Getestet wurde ein vierfaktorielles CFA-Modell, in dem die Items, wie zuvor beschrieben, entsprechend ihrer höchsten Ladung genau einem der vier Faktoren zugeordnet wurden. Der Fit des Modells fiel auf Grund der nicht zugelassenen Nebenladungen erwartungsgemäß schlechter aus als der des EFA-Modells: $\chi^2 = 743.46$, $df = 399$, $p < .001$, CFI = .81, RMSEA = .08, 90% CI = [.07, .08] (MLR-Schätzer). Aus diesem Grund wurde zusätzlich ein Exploratives Strukturgleichungsmodell (ESEM-Modell) geschätzt, in dem sämtliche Nebenladungen zugelassen wurden. Der Fit des ESEM-Modells fiel besser, aber weiterhin nicht zufriedenstellend aus: $\chi^2 = 538.95$, $df = 321$, $p < .001$, CFI = .88, RMSEA = .07, 90% CI = [.06, .08] (MLR-Schätzer). Allerdings zeigte sich bei der Betrachtung des Ladungsmusters des ESEM-Modells, dass die Skalenzuordnung bei 28 der insgesamt 30 Items repliziert werden konnte, was heißt, dass 28 Items im ESEM-Modell die höchste Ladung auf demselben Faktor zeigten wie im EFA-Modell.

Letztendlich ist anzumerken, dass von den acht Items, die als Grundlage für die Kontingenzen der AcquA-PU Aufgaben PU1, PU2 und PU3 verwendet wurden, ein Item im Laufe der Itemselektion eliminiert wurde (Aufgabe PU3).

Verträglichkeit. Die ML-EFAs der Verträglichkeits-Items wurden analog zu jenen der Gewissenhaftigkeits-Items durchgeführt und ebenfalls auf Grund der univariaten Verteilungen der Items unter Verwendung des MLR-Schätzers. Eine Parallelanalyse aller 31 Items deutete auf sechs zu Grunde liegende Faktoren hin (Vergleich empirische Eigenwerte mit dem .95-

Quantil der Eigenwerte der 5000 Zufallsdaten) und der Screeplot zeigte einen uneindeutigen Eigenwertverlauf. Das sechsfaktorielle EFA-Modell zeigte einen nur teilweise guten Fit (χ^2 = 461.75, *df* = 294, $p < .001$, CFI = .92, RMSEA = .05, 90% CI = [.04, .06]), der aber signifikant besser ausfiel als der Fit des fünffaktoriellen Modells ($\tilde{T}_d$= 121.71, *df* = 26, $p < .001$, ΔCFI = .04, ΔRMSEA = .01). Anschließend wurde das rotierte Ladungsmuster betrachtet, wobei im Falle der Verträglichkeits-Items die oblique Varimax-Rotation[12] verwendet wurde. Hierbei zeigte sich, dass die sechsfaktorielle Lösung zu einem fünften und sechsten Faktor führt, auf denen jeweils nur zwei Items Primärladungen ≥ |.30| aufweisen. In beiden Fällen handelte es sich zudem um inhaltlich weitestgehend redundante Items: „Ich bin leicht zufriedenzustellen" (Item 11, Ladung: -.75) und „Ich bin schwer zufriedenzustellen" (Item 17, Ladung: .87) sowie „Ich vertraue anderen" (Item 14, Ladung: .62) und „Ich vertraue auf das, was andere sagen" (Item 16, Ladung: .78). Zur Vermeidung von Methodenfaktoren wurden die Items 17 und 16 eliminiert, wobei dies iterativ in der genannten Reihenfolge erfolgte und zwischendurch die Dimensionalität und das Ladungsmuster erneut kontrolliert wurden.

Die Parallelanalyse auf Basis des reduzierten Pools mit 29 Items ergab fünf zu extrahierende Faktoren (vgl. Tabelle D7 in Anhang D). Auch der Screeplot (vgl. Abbildung D2 in Anhang D) war mit dieser Faktoranzahl zu vereinbaren. Das fünffaktorielle EFA-Modell zeigte mit χ^2 = 431.18, *df* = 271, $p < .001$, CFI = .92, RMSEA = .05, 90% CI = [.04, .06] erneut nicht durchweg einen guten Fit. Ein Vergleich mit dem sechsfaktoriellen Modell war nicht möglich, da dieses nicht konvergierte. Die inhaltliche Betrachtung des Ladungsmusters ergab inhaltlich sinnvoll interpretierbare Faktoren.

Die nach Extraktion von fünf Faktoren durchgeführte Itemselektion erfolgte analog zu der Selektion der Gewissenhaftigkeits-Items. In einem iterativen Prozess wurden zunächst vier Items eliminiert, die auf keinem Faktor eine Primärladung ≥ |.30| aufwiesen. Anschließend wurde ein weiteres Item mit einer Sekundärladung ≥ |.30| eliminiert, dessen Elimination zu einer Verbesserung im Fit führte. Auch wenn fünf weitere Sekundärladungen ≥ |.30| vorlagen, wurde keine weitere Eliminationen vorgenommen, da die Kriterien CFI ≥ .95 und RMSEA ≤ .05 erreicht wurden. Die vollständige Vermeidung von Sekundärladungen wurde für das Ziel der vorliegenden Studie, das nicht in der Bildung eines Instruments für Verträglichkeit, sondern in der Schätzung des Zusammenhangs zu PU bestand, als nicht relevant erachtet. Daher wurden

[12] Erneut wurden neben der obliquen Varimax-Rotation auch die Geomin-Rotation sowie die Quartimin-Rotation exploriert. Die drei Rotationen kamen im Rahmen der Itemselektion zu leicht abweichenden Ergebnissen (eliminierte Items, Zuordnung der einzelnen Items zu den Faktoren). Insgesamt zeigte die oblique Varimax-Rotation die besten Ergebnisse, insbesondere bezüglich der Interpretation der Faktoren.

die fünf Items mit den Doppelladungen beiden Faktoren, auf denen sie Ladungen ≥ |.30| aufwiesen, zugeordnet (vgl. Tabelle D8 in Anhang D). Eine inhaltliche Betrachtung aller 24 zugeordneten Items deutet darauf hin, dass Faktor 1 das Interesse an den Emotionen und Sorgen anderer, Faktor 2 Hilfsbereitschaft, Faktor 3 Respekt gegenüber anderen, Faktor 4 (fehlende) Konfliktbereitschaft und Faktor 5 (fehlende) Freundlichkeit widerspiegelt.

5.3.7.4 Zusammenhang PU-Parcels und Skalen der IPIP- Items. Schließlich wurden die Zusammenhänge zwischen den faktorenanalytisch gebildeten Gewissenhaftigkeits-Skalen und Verträglichkeits-Skalen sowie den hypothesengeleitet gebildeten PU-Parcels geschätzt. Deskriptive Statistiken und Reliabilitätsschätzungen der Skalen sowie der Parcels sind in Tabelle 5.12 dargestellt (für die PU-Parcels des dichotomen Scorings siehe Tabelle C8 in Anhang C). Auffällig sind hier insbesondere die geringen Reliabilitäten einiger PU-Parcels. Zudem sollte bei der Interpretation der folgenden Ergebnisse berücksichtigt werden, dass nur wenige Items in die Bildung der PU-Parcels des Faktors Verträglichkeit eingeflossen sind, da nur eine AcquA-PU Aufgabe dieses Faktors verwendet werden konnte (PU6).

Die Zusammenhänge zwischen den Skalen und den PU-Parcels befinden sich in Tabelle 5.13 (für die Ergebnisse unter Verwendung des dichotomen Scorings siehe Tabelle C9 in Anhang C). Bei der Interpretation dieser Zusammenhänge ist abermals zu beachten, dass bei PU auf Grund der standardisierten Distanzen höhere Werte eine schlechtere Leistung anzeigen.

Unter Verwendung des standardisierten Distanzscorings resultierten für die Skalen der Gewissenhaftigkeit maximal geringe und nicht signifikante Zusammenhänge zu den PU-Parcels. Unter Verwendung des dichotomen Scorings ergab sich ein ähnliches Ergebnismuster mit auf Grund der Scoring-Methode erwartungsgemäß gegenteiligen Vorzeichen der Korrelationen (vgl. Tabelle C9, Anhang C). Hier zeigte sich lediglich eine signifikante Korrelation zwischen dem PU-Parcel Pc_high sowie Skala 1 (Einhaltung von Regeln und Verpflichtungen), die mit r = -.15 zum einen nicht hoch ausfiel und zum anderen im Hinblick auf einen möglichen Effekt der Ähnlichkeit zwischen Testperson und Zielperson eine nicht erwartete Richtung aufwies.

Tabelle 5.12

*Itemanzahl (*k*), deskriptive Statistiken sowie Reliabilitätsschätzungen (ω_H; Kelley & Pornprasertmanit, 2016) der faktorenanalytisch gebildeten Gewissenhaftigkeits-Skalen und Verträglichkeits-Skalen sowie der hypothesengeleitet gebildeten Personality Understanding (PU)-Parcels (standardisiertes Distanzscoring)*

Skala/Parcel	*k*	*M*	*SD*	Schiefe[a]	Kurtosis[b]	ω_H [95% CI]
Gewissenhaftigkeit						
F1	12	4.73	0.74	-0.54	-0.30	.87 [.84, .89]
F2	6	4.22	0.96	-0.25	-0.63	.86 [.83, .89]
F3	6	3.41	1.12	0.07	-0.54	.87 [.83, .91]
F4	6	3.90	1.20	-0.12	-0.92	.90 [.87, .92]
Verträglichkeit						
F1	7	5.15	0.80	-1.51	2.27	.90 [.87, .92]
F2	6	4.93	0.71	-0.96	1.28	.78 [.72, .84]
F3	6	4.97	0.74	-1.27	2.42	.77 [.68, .83]
F4	6	4.34	0.85	-0.54	0.09	.73 [.66, .79]
F5	4	4.27	0.91	-0.20	0.00	.57 [.46, .66]
Personality Understanding						
Pc_high	13	0.46	0.16	0.60	0.21	.56 [.39, .65][c]
Pc_low	14	0.51	0.23	1.08	0.91	.72 [.63, .78][c]
Pa_high	5	0.52	0.25	0.79	0.72	.42 [.32, .72][c]
Pa_low	2	0.90	0.28	0.07	2.11	- [d]

Anmerkungen. N = 204. CI = Konfidenzintervall. Aufbau PU-Parcelname: P = PU, c = Gewissenhaftigkeit, a = Verträglichkeit, high = Items des Parcels beziehen sich auf Zielpersonen mit hoher Ausprägung der Persönlichkeitseigenschaft, low = Items des Parcels beziehen sich auf Zielpersonen mit niedriger Ausprägung der Persönlichkeitseigenschaft.

[a] *SE* = 0.17. [b] *SE* = 0.34 [c] Bei Schätzung des Konfidenzintervalls traten wiederholt Schätzprobleme (insb. keine Konvergenz, negative Varianzen) auf. [d] Reliabilität wurde auf Grund der Itemanzahl nicht geschätzt.

Tabelle 5.13

Korrelationen zwischen den hypothesengeleitet gebildeten Personality Understanding (PU)-Parcels (standardisiertes Distanzscoring) und den faktorenanalytisch gebildeten Gewissenhaftigkeits-Skalen und Verträglichkeits-Skalen

	Gewissenhaftigkeit				Verträglichkeit			
PU-Parcel	F1	F2	F3	F4	F1	F2	F3	F4
Pc_high	.12	.08	.09	.06	-.06	.01	.12	.01
Pc_low	.05	.03	.06	.12	-.04	-.09	.01	-.01
Pa_high	-.02	-.07	.06	-.08	-.25***	-.19**	-.08	-.08
Pa_low	-.14	-.05	-.04	-.11	-.20**	-.02	-.03	-.05

Anmerkungen. N = 204. Aufbau Parcelname: P = PU, c = Gewissenhaftigkeit, a = Verträglichkeit, high = Items des Parcels beziehen sich auf Zielpersonen mit hoher Ausprägung der Persönlichkeitseigenschaft, low = Items des Parcels beziehen sich auf Zielpersonen mit niedriger Ausprägung der Persönlichkeitseigenschaft. Für eine bessere Übersicht über die Ergebnisse werden keine Konfidenzintervalle der Korrelationen berichtet.

**$p < .01$. ***$p < .001$.

Für die Verträglichkeits-Skalen ergaben sich maximal geringe und nicht signifikante Zusammenhänge mit den PU-Parcels der Gewissenhaftigkeit. Bei den PU-Parcels, die hoch und niedrig verträglichen Zielpersonen zugeordnet werden können, zeigten sich hingegen drei signifikante Zusammenhänge mit den Verträglichkeits-Skalen 1 (Interesse an den Emotionen und Sorgen anderer) und 2 (Hilfsbereitschaft). Die IPIP-Items, die zur Konstruktion der Kontingenzen dieser Zielpersonen verwendet wurden („Ich schätze Kooperation mehr als Konkurrenz“, „Ich suche den Konflikt“), sind allerdings nicht in den Verträglichkeits-Skalen 1 und 2 enthalten. Für alle drei signifikanten Korrelationen gilt, dass diese darauf hindeuten, dass eine höhere Ausprägung auf der Verträglichkeits-Skala mit einer besseren Leistung bei der Einschätzung der hoch beziehungsweise niedrig verträglichen Zielperson einhergeht. Die Ergebnisse unter Verwendung des dichotomen Scorings zeigten diese Zusammenhänge nicht (vgl. Tabelle C9, Anhang C). Hier ist allerdings zu beachten, dass die PU-Parcels der Verträglichkeit nur aus einem beziehungsweise zwei Items bestehen.

5.4 Diskussion

In der zweiten Studie wurden die vorhandenen AcquA-Aufgaben (3D-Format) zur Erfassung von PU vertont und überarbeitet, neue AcquA-PU Aufgaben konstruiert sowie weitere Validitätsevidenz gesammelt. Zwei der drei neu konstruierten Aufgaben zeigten im Messmodell sehr geringe Ladungen, sodass sie in den weiteren Analysen nicht berücksichtigt wurden. Auch wenn die Eliminierung beider Aufgaben zu einer Verbesserung der Reliabilität führte, fiel diese auf Basis der vier verbliebenden Aufgaben zu gering aus. Die latenten Korrelationen zwischen PU und den Big Five lieferten zusätzliche Evidenz für diskriminante Validität, während der latente Zusammenhang zwischen PU und schlussfolgerndem Denken insbesondere beim dichotomen Scoring deutlich geringer ausfiel als erwartet. Bei der Untersuchung des Zusammenhangs zwischen PU und der Persönlichkeit der Testperson, wenn Zielperson und Testpersonen eine ähnliche Ausprägung auf der betrachteten Persönlichkeitseigenschaft aufweisen, zeigten sich spezifische Zusammenhänge zwischen Parcels einer AcquA-PU Aufgabe, in der verträgliches Verhalten modelliert wurde, und zwei neu gebildeten Skalen zur Erfassung von Verträglichkeit.

Im Rahmen der Itemselektion sowie im Messmodell wurden unter Verwendung beider Scoring-Methoden Probleme der neu konstruierten Aufgaben PU4 und PU5 deutlich. Auch wenn eine nachträgliche Erklärung dieser Probleme nicht eindeutig möglich ist, erscheint es am wahrscheinlichsten, dass die Hauptursache für die schlechten psychometrischen Ergebnisse in den relevanten Kontingenzen beider Aufgaben liegt. Die Kontingenzen sind das Kernelement der AcquA-Aufgaben (Schulze & Roberts, 2015) und werden bei PU zur Modellierung der Persönlichkeit der Zielpersonen in der Aneignungsphase sowie zur Bewertung der Testpersonenantworten in der Anwendungsphase verwendet. Es ist daher denkbar, dass unpassende oder missverständliche Kontingenzen eine reliable und valide Erfassung von PU erschwert oder verhindert haben. Eine Überarbeitung der Kontingenzen sollte daher in Erwägung gezogen werden. Hierbei und für künftige Aufgabenkonstruktionen könnten (qualitative) Pretests der Kontingenzen von Vorteil sein. Durch diese ließe sich überprüfen, ob die Kontingenzen wie intendiert verstanden werden und missverständliche Kontingenzen könnten von vornherein ausgeschlossen werden.

Nicht nur auf Grund der schlechten psychometrischen Ergebnisse der Aufgaben PU4 und PU5 musste in der zweiten Studie eine sehr große Anzahl an Items eliminiert werden. Während unter Verwendung der standardisierten Distanzen die Items aller Aufgaben um fast die Hälfte reduziert wurden, blieben beim dichotomen Scoring nach der Itemselektion nur noch

etwas über ein Drittel der Items übrig. Grund für die Diskrepanz beider Scoring-Methoden war das zusätzliche Selektionskriterium, das beim dichotomen Scoring angewendet wurde (u.a. Elimination von Items mit Mittelwerten > .95) und aufgrund dessen mehr als ein Drittel der Items eliminiert wurden. Dies verdeutlicht die bereits in der vorherigen Studie angesprochene Problematik, dass die AcquA-PU Aufgaben aktuell sehr leicht sind. Zudem zeigten beim dichotomen Scoring trotz des zusätzlichen Selektionskriterium alle vier Aufgaben des finalen Messmodells Mittelwerte > .50 und somit Werte im eher leichten Schwierigkeitsbereich. Und auch beim standardisierten Distanzscoring deuten die Ergebnisse zur Schiefe der Aufgaben-Parcels darauf hin, dass diese eine eher geringe Schwierigkeit aufweisen. In kommenden Aufgabenkonstruktionen sollte daher die gezielte Konstruktion schwierigerer Aufgaben im Mittelpunkt stehen.

Der Fit des Messmodells für PU auf Basis der vier verbliebenen Aufgaben, von denen drei dem Big Five-Faktor Gewissenhaftigkeit zuzuordnen sind und eine dem Faktor Verträglichkeit, liefert klare Evidenz für faktorielle Validität der AcquA-PU Aufgaben. Beim dichotomen Scoring zeigte die Verträglichkeits-Aufgabe zwar eine geringere Ladung als die Gewissenhaftigkeits-Aufgaben, dies ist allerdings eher auf die bei dieser Aufgabe geringere Anzahl an selektierten Items zurückzuführen und nicht auf den anderen Big Five-Faktor. Dass die Reliabilität auf Basis der vier Aufgaben nicht zufriedenstellend sowie geringer als in Studie 1 ausfiel, ist der geringen Anzahl an Aufgaben zuzuschreiben. Zudem fiel die Reliabilität beim dichotomen Scoring noch geringer aus als beim standardisierten Distanzscoring. Dies ist einerseits mit der dort deutlich höheren Anzahl an eliminierten Items zu vereinbaren und stimmt andererseits mit den Ergebnissen der Vorstudie – wo beim dichotomen Scoring allerdings ähnlich viele Items eliminiert wurden wie beim standardisierten Distanzscoring – überein. In jedem Fall ist es in kommenden Studien zur Steigerung der Reliabilität erforderlich, die Aufgaben PU4 und PU5 zu überarbeiten oder neue Aufgaben zu konstruieren.

Einschränkend muss im Zusammenhang mit den überarbeiteten und neu konstruierten Aufgaben angemerkt werden, dass unklar ist, welchen Effekt die Vertonung der Dialoge hatte. Während in der ersten Studie die Übertragung vom 2D-Format auf das noch nicht vertonte 3D-Format systematisch variiert wurde, wurden in der vorliegenden Studie alle Aufgaben vertont. Ein Methodeneffekt der Vertonung kann somit nicht gänzlich ausgeschlossen werden. Da aber bereits die Übertragung auf das neue 3D-Format mit keinem substantiellen Methodeneffekt einherging, sollte Ähnliches für die Vertonung der Aufgaben angenommen werden können. Zudem konnte für das konzeptuell verwandte Konstrukt SU bereits gezeigt werden, dass unterschiedliche Präsentationsmethoden der sozialen Information (geschriebener Text,

gesprochener Text, Bilder, Videos ohne Ton; Weis et al., 2006) die Erfassung von SU gleichermaßen gut erlauben (Baumgarten, 2015; Seidel, 2007; Weis, 2008).

Der Zusammenhang zwischen PU und schlussfolgerndem Denken fiel unter Verwendung beider Scoring-Methoden, aber insbesondere beim dichotomen Scoring, deutlich geringer aus als erwartet. Evidenz für konvergente Validität konnte somit nicht gesammelt werden. Bei der Interpretation des Ergebnisses muss jedoch die eben thematisierte geringe Reliabilität der AcquA-PU Aufgaben berücksichtigt werden. Der geschätzte Zusammenhang zwischen beiden Konstrukten ist daher mit einer entsprechenden Unsicherheit verbunden, die auch bei Betrachtung der Konfidenzintervalle deutlich wird. Wünschenswert wäre daher eine erneute Untersuchung des Zusammenhangs unter Verwendung einer reliableren Zusammenstellung von AcquA-PU Aufgaben. Eine weitere mögliche Erklärung für den geringen Zusammenhang wäre, dass schlussfolgerndes Denken (beziehungsweise logisches Schlussfolgern) nicht die zentrale kognitive Anforderung der Aufgaben darstellt oder zumindest weitere Fähigkeiten in einem größeren Umfang als vermutet bei der Aufgabenlösung involviert sind. Bei Hellwig et al. (2020) sowie Schulze und Jobmann (2016) resultierte ein mittlerer beziehungsweise großer latenter Zusammenhang zwischen EU und Merkfähigkeit. Diese Fähigkeit spielt auch bei der Operationalisierung von PU keine unwichtige Rolle. Merkfähigkeit wird durch das AcquA-Testdesign automatisch mit angesprochen, da sich die Testpersonen in der Aneignungsphase die typischen Reaktionen der Zielpersonen einprägen, kurzfristig behalten und in der Anwendungsphase schließlich erinnern und anwenden müssen (Hellwig et al., 2020). Konzeptuell stellt das Einprägen und Erinnern von Reaktionen und Verhaltensweisen der Zielperson zudem eine Voraussetzung für PU dar, da ansonsten die Grundlage für das logische Schlussfolgern über deren Persönlichkeit fehlt. Inwieweit Merkfähigkeit die Leistung bei den AcquA-PU Aufgaben mitbestimmt, muss letztendlich empirisch geklärt werden.

Die maximal geringen und überwiegend nicht signifikanten Zusammenhänge zwischen PU und den Big Five-Faktoren sind insgesamt vergleichbar mit den Ergebnissen der ersten Studie und fielen nur in einzelnen Modellen deskriptiv etwas höher aus. Die Ergebnisse liefern somit weitere Evidenz für diskriminante Validität der AcquA-PU Aufgaben.

Bezüglich der Untersuchung eines möglichen Einflusses der Ähnlichkeit zwischen Testperson und Zielperson muss zunächst einmal festgehalten werden, dass Studie 2 weitere Hinweise auf eine geringe Stabilität der AB5C-Skalen geliefert hat. Die in der vorherigen Studie durchgeführte Skalenbildung konnte überwiegend nicht repliziert werden und zudem traten die dort beobachteten Probleme bei der Anwendung des AB5C-Algorithmus von Hofstee

et al. (1992) erneut auf (z.B. ähnlich hohe Nebenladungen auf mehreren Faktoren). Die Ergebnisse weisen in Übereinstimmung mit Saucier und Ostendorf (1999) somit darauf hin, dass eine Skalenbildung mittels AB5C-Algorithmus auf Basis einer deutlich größeren Stichprobe vorgenommen werden muss, da sonst keine stabilen Ergebnisse erzielt werden können. Die AB5C-Skalen wurden daher verworfen und neue Skalen auf Basis von Faktorenanalysen gebildet, die eine inhaltlich interpretierbare Struktur zeigten. Hierbei ist zu bedenken, dass zur Erfassung von Gewissenhaftigkeit und Verträglichkeit jeweils die Items von lediglich drei der insgesamt neun AB5C-Skalen von Goldberg (1999) eingesetzt wurden. Beide Faktoren wurden durch die verwendeten Items somit nicht in ihrer gesamten inhaltlichen Breite abgebildet, sodass sich auch mögliche Facetten der Faktoren nur eingeschränkt in den Ergebnissen zeigen konnten. Im Folgenden soll trotzdem ein Vergleich der faktorenanalytisch gebildeten Skalen mit den Facetten der Big Five, wie sie im NEO-PI-R erfasst werden (McCrae & John, 1992; Ostendorf & Angleitner, 2004), sowie Ergebnissen weiterer Studien vorgenommen werden, um die Interpretierbarkeit der neu gebildeten Skalen und damit deren Nutzen besser abschätzen zu können.

Die erste der insgesamt vier neu gebildeten Gewissenhaftigkeits-Skalen (C-Skalen) repräsentiert am ehesten die Einhaltung von Regeln und Verpflichtungen. Damit weist sie große Gemeinsamkeiten mit der Facette Pflichtbewusstsein aus dem NEO-PI-R auf, die ebenfalls die Tendenz zum Erfüllen von insbesondere ethischen Prinzipien und Verpflichtungen umfasst (Ostendorf & Angleitner, 2004). Kleinere Überschneidungen liegen zudem zur NEO-PI-R-Facette Selbstdisziplin vor, die unter anderem die Neigung beschreibt, begonnene Aufgaben zu beenden (Ostendorf & Angleitner, 2004). C-Skala 2 (planvolles Handeln) ist eine sehr spezifische Skala, die sich in der Form im NEO-PI-R lediglich als ein kleiner Bestandteil der Facette Ordnungsliebe wiederfindet (Ostendorf & Angleitner, 2004). Sehr große Überschneidungen zeigen sich wiederum mit der von MacCann et al. (2009) identifizierten Facette Task Planning, die teilweise dieselben Items wie C-Skala 2 enthält und bei der ebenfalls das Befolgen von Plänen und Routinen im Zentrum steht. Auch C-Skala 3 (Prokrastination) zeigt sich in ähnlicher Form bei MacCann et al. (2009) unter der Bezeichnung Procrastination Refrainment. In Bezug auf das NEO-PI-R finden sich Überschneidungen mit der Facette Selbstdisziplin, die neben dem oben genannten Aspekt auch das Aufschieben von Arbeit umfasst (Ostendorf & Angleitner, 2004). Bei C-Skala 4 (Vorliebe für Ordnung) finden sich klare Überschneidungen mit der NEO-PI-R-Facette Ordnungsliebe (Ostendorf & Angleitner, 2004) sowie mit der Facette Tidiness bei MacCann et al. (2009). In allen drei Skalen beziehungsweise Facetten steht die Tendenz, Ordnung zu halten und zu bevorzugen, im

Mittelpunkt. C-Skala 4 weist zudem Gemeinsamkeiten mit einem von zwei zentralen Aspekten der Gewissenhaftigkeit auf, die DeYoung et al. (2007) in einer Studie identifizierten und als Orderliness bezeichneten.

Bei den neu gebildeten Verträglichkeits-Skalen (A-Skalen) können Gemeinsamkeiten von A-Skala 1 (Interesse an den Emotionen und Sorgen anderer) und der Facette Gutherzigkeit des NEO-PI-R gefunden werden. Gutherzigkeit umfasst unter anderem die Orientierung am Wohl und den Bedürfnissen anderer Personen, berücksichtigt aber nicht das Interesse an den Emotionen anderer Personen (Ostendorf & Angleitner, 2004). Dieses ist dafür integraler Bestandteil der Skala Compassion, die DeYoung et al. (2007) als einen von zwei zentralen Aspekten der Verträglichkeit identifiziert haben. Compassion beschreibt die Tendenz, sich anderen emotional zugehörig zu fühlen sowie das Interesse an den Emotionen anderer (DeYoung et al., 2007). Somit liegen klare Überschneidungen von A-Skala 1 und der Skala Compassion vor, auch da beide Skalen teilweise identische Items beinhalten (vgl. DeYoung et al., 2007). Die neu gebildete A-Skala 2 (Hilfsbereitschaft) deckt den restlichen Teil der Facette Compassion ab, die auch die Neigung beinhaltet, anderen Personen zu helfen (DeYoung et al., 2007). A-Skala 2 weist in diesem Punkt zudem Gemeinsamkeiten mit der Facette Altruismus aus dem NEO-PI-R auf (Ostendorf & Angleitner, 2004). A-Skala 4 (Konfliktbereitschaft) weist die größten Gemeinsamkeiten mit der NEO-PI-R-Facette Entgegenkommen auf, in der es ebenfalls um das Verhalten bei Konflikten mit anderen Personen geht (Ostendorf & Angleitner, 2004). Bei den A-Skalen 3 (Respekt gegenüber anderen) und 5 (Freundlichkeit) sind die Gemeinsamkeiten mit anderen Skalen und Facetten weniger eindeutig. Für A-Skala 3 lassen sich zumindest geringfügige Gemeinsamkeiten mit Items der NEO-PI-R-Facette Altruismus finden (vgl. Ostendorf & Angleitner, 2004). Zudem finden sich Überschneidungen mit der Skala Respectfulness aus dem BFI-2, die den freundlichen und respektvollen Umgang mit anderen widerspiegelt (Soto & John, 2017).

Insgesamt zeigt der vorgenommene Vergleich, dass in der vorliegenden Studie, trotz stark eingeschränkter Itemgrundlage, weitestgehend inhaltlich sinnvolle Skalen identifiziert wurden. Abgesehen von der A-Skala 5, die nur vier Items umfasst, resultierten zudem für alle Skalen gute Reliabilitätsschätzungen. Die durchgeführte Kreuzvalidierung der C-Skalen, bei der sich kein zufriedenstellender Fit des CFA-Modells zeigte, lässt zwar Zweifel an der Stabilität der Skalenbildung zu – anders als bei den AB5C-Skalen resultierte jedoch sowohl in der Ausgangsstichprobe als auch in der Validierungsstichprobe für fast alle Items dieselbe Skalenzuordnung, sodass zumindest eine gewisse Stabilität angenommen werden kann. Für die A-Skalen steht eine Kreuzvalidierung noch aus.

Zwischen den C-Skalen und den hypothesengeleitet gebildeten PU-Parcels zeigten sich keine Zusammenhänge, die auf einen möglichen Effekt der Ähnlichkeit zwischen Testperson und Zielperson hindeuten würden. Interessant sind allerdings die beim standardisierten Distanzscoring gefundenen Zusammenhänge zwischen den PU-Parcels, die sich auf hoch und niedrig verträgliche Zielpersonen beziehen, sowie den A-Skalen 1 und 2, die Interesse an und Bemühen um andere Personen widerspiegeln.[13] Diese Zusammenhänge sind nicht auf die mit den beiden A-Skalen erfassten interpersonalen Persönlichkeitseigenschaften an sich zurückzuführen, da sich die Zusammenhänge nur für die PU-Parcels der hoch und niedrig verträglichen Zielpersonen gezeigt haben, nicht aber für die PU-Parcels der hoch und niedrig gewissenhaften Zielpersonen. Dass sich die Zusammenhänge spezifisch bei den PU-Parcels und Skalen der Verträglichkeit gezeigt haben, passt zudem zu den Ergebnissen von Thielmann et al. (2020). Den Autor:innen zufolge nehmen Beurteiler:innen insbesondere für die Persönlichkeitsfaktoren Offenheit für Erfahrungen und Verträglichkeit, aber eben nicht für die Gewissenhaftigkeit, Ähnlichkeit zwischen sich selbst und den Zielpersonen an. Und nur wenn Ähnlichkeit zur Zielperson angenommen wird, kann die tatsächliche Ähnlichkeit eine bessere Teilleistung zur Folge haben (vgl. Paunonen & Hong, 2013; Vogt & Colvin, 2003). Nach Thielmann et al. (2020) ist der entscheidende Faktor, dass Beurteiler:innen Ähnlichkeit zur eigenen Person nur bei Persönlichkeitseigenschaften annehmen, die persönliche Werte, Fairness und Moral betreffen (vgl. auch Thielmann et al., 2022). Genau diese zwei Aspekte sind Bestandteil der beiden PU-Parcels des Faktors Verträglichkeit. So geht es im ersten Parcel (Pa_high) um die Einschätzung, ob die Zielperson eher kooperiert oder mit anderen Personen konkurriert, und im zweiten Parcel (Pa_low) um die Einschätzung von Konfliktverhalten der Zielperson. Insgesamt scheint daher eine Erklärung der gefundenen Zusammenhänge über die tatsächliche Ähnlichkeit am naheliegendsten. Allerdings muss diese Schlussfolgerung aus mehreren Gründen unter Vorbehalt gemacht werden. Zum einen weisen die beiden hypothesengeleiteten PU-Parcels der Verträglichkeit inhaltlich die größten Überschneidungen mit den A-Skalen 4 und 5 und nicht mit den A-Skalen 1 und 2 auf. Zum anderen resultierten für die A-Skala 1 und die Parcels Pa_high und Pa_low Zusammenhänge in derselben Richtung. Dies ist allerdings nicht mit einem Effekt der tatsächlichen Ähnlichkeit zu vereinbaren, da beim ersten Parcel hoch verträgliche Testpersonen Ähnlichkeit zur Zielperson aufweisen und beim zweiten Parcel niedrig verträgliche Testpersonen. Beachtet werden muss wiederum, dass die

[13] Anzumerken ist hierbei, dass sich die beiden A-Skalen 1 und 2 auf Grund von hohen Sekundärladungen zwei Items teilen (vgl. Tabelle D8 in Anhang D).

Reliabilitäten der PU-Parcels sehr gering ausfielen beziehungsweise die Reliabilität im Fall des Parcels Pa_low auf Grund der wenigen Items nicht geschätzt wurde. Insbesondere der Zusammenhang mit Parcel Pa_low ist daher kaum aussagekräftig. Die Hauptursache der geringen Reliabilitäten liegt darin, dass – anders als ursprünglich geplant – auf Grund der schlechten Ergebnisse der Aufgaben PU4 und PU5 für die Analysen nur eine PU-Aufgabe mit Fokus auf Verträglichkeit genutzt werden konnte. Die Zusammenhänge könnten somit auch ein Spezifikum der Aufgabe PU6 darstellen. Dass sich die Zusammenhänge beim standardisierten Distanzscoring zeigten, nicht aber beim dichotomen Scoring, stellt hingegen keinen direkten Widerspruch dar. So bestand das Parcel Pa_high beim dichotomen Scoring aus lediglich einem Item, was für eine angemessene Interpretation nicht ausreicht. Insgesamt muss die Frage nach einem möglichen Effekt der Ähnlichkeit zwischen Zielperson und Testperson also als weiterhin nicht eindeutig geklärt betrachtet werden.

Alles in allem konnte die zweite Studie die bereits vorhandene Evidenz für faktorielle und diskriminante Validität der AcquA-PU Aufgaben stützen und auf die Aufgaben im vertonten 3D-Format ausweiten. Der Zusammenhang zwischen PU und logischem Schlussfolgern wäre ein wertvoller Hinweis auf konvergente Validität gewesen, der allerdings und möglicherweise auf Grund der zu geringen Reliabilität der AcquA-PU Aufgaben nicht erbracht werden konnte. Im Fokus folgender Untersuchungen sollte daher weiterhin die Untersuchung der Konstruktvalidität – und hier insbesondere der konvergenten Validität – der AcquA-PU Aufgaben stehen. Auf Grund der Ergebnisse im Zusammenhang mit den Aufgaben PU4 und PU5 ist eine hierfür notwendige Voraussetzung die Überarbeitung oder Erweiterung des vorhandenen Aufgabenpools, um eine reliable Erfassung von PU zu ermöglichen. Darüber hinaus weisen die Ergebnisse der zweiten Studie in Übereinstimmung mit Studie 1 darauf hin, dass die gezielte Konstruktion schwieriger Aufgaben angestrebt werden sollte.

6. Studie 3

6.1 Ziele

Da die bisherigen Ergebnisse zur Konstruktvalidität der AcquA-PU Aufgaben heterogen ausfielen, bestand das erste Ziel in der Sammlung weiterer Evidenz für insbesondere konvergente Validität. Wie in Abschnitt 3.1.1 ausführlich erläutert wurde, handelt es sich bei PU und SU – ähnlich wie beim Vergleich zwischen PU und EU – um konzeptuell sehr ähnliche, aber trotzdem inhaltlich voneinander abgrenzbare Konstrukte. Die Untersuchung des Zusammenhangs der AcquA-PU Aufgaben mit einem Maß für SU sollte daher weitere Hinweise auf die Konstruktvalidität liefern. Erwartet wurde auf Grund der hohen konzeptuellen Ähnlichkeit beider Konstrukte bei gleichzeitiger inhaltlicher Unterscheidbarkeit auf latenter Ebene ein mittlerer positiver Zusammenhang. Darüber hinaus sollte im Rahmen des ersten Ziels auch der Zusammenhang zwischen den AcquA-PU Aufgaben sowie der Akkuratheit von Persönlichkeitsbeurteilungen, unter Verwendung der sehr häufig bestimmten Selbst-Fremd Übereinstimmung (vgl. Funder, 2012), untersucht werden. Die Bestimmung der Akkuratheit kann als klassische Operationalisierung einer Fähigkeit zu akkuraten Beurteilungen der Persönlichkeit anderer Personen aufgefasst werden und wird von einigen Autor:innen auch so interpretiert (z.B. Biesanz, 2010; Cronbach, 1955; Funder, 1999; Hall et al., 2018). Daher sollte der Zusammenhang zwischen beiden Methoden weitere konvergente Validitätsevidenz liefern. Allerdings ist der klassische Ansatz der Selbst-Fremd Übereinstimmung problembehaftet (vgl. Abschnitt 3.2) und unterscheidet sich substantiell von dem Ansatz auf Basis des AcquA-Testdesigns. Daher wurde zwischen beiden Methoden auf latenter Ebene ein mittlerer positiver, aber kein großer Zusammenhang erwartet.

Das zweite Ziel ergab sich aus den Ergebnissen der vorherigen Studie, wo insbesondere der Zusammenhang zwischen PU und logischem Schlussfolgern nicht erwartungskonform ausfiel. Da der Zusammenhang im Hinblick auf die Konstruktvalidität der AcquA-PU Aufgaben allerdings zentral und das Ergebnis der vorherigen Studie auf Grund der geringen Reliabilität der AcquA-PU Aufgaben mit einer gewissen Unsicherheit verbunden ist, sollte der Zusammenhang in der dritten Studie erneut betrachtet werden. Im Rahmen dessen sollte zudem untersucht werden, ob die AcquA-PU Aufgaben anstelle von logischem Schlussfolgern eine andere zentrale Fähigkeit oder eine weitere Fähigkeit in einem größeren Ausmaß als erwartet erfordern. Hier war die Rolle der Merkfähigkeit von Interesse. Wie in der Diskussion der vorherigen Studie erläutert, ist bei PU das Einprägen und Erinnern typischer Verhaltensweisen

der Zielperson sowohl aus theoretischer Sicht als auch bei der Operationalisierung mittels AcquA-Testdesign von Bedeutung. So kann diese Fähigkeit als Voraussetzung für PU angesehen werden und wird zudem durch das Zwei-Phasen-Design des AcquA-Testdesigns mit angesprochen (Hellwig et al., 2020). Die Unterschiede zwischen theoretischer Ebene und Operationalisierungsebene bestehen allerdings zum einen darin, dass auf theoretischer Ebene das Einprägen und Erinnern zwar als Voraussetzung, aber nicht als Bestandteil von PU angesehen werden kann, bei den AcquA-Aufgaben aber zwangsläufig miterfasst wird. Zum anderen wird bei PU das Erinnern über verschieden lange und auch längere Zeiträume hinweg benötigt – auf Ebene der Operationalisierung ist das Behalten der Information hingegen nur über einen relativ kurzen Zeitraum von wenigen Minuten gefordert (Hellwig et al., 2020). Beim AcquA-Testdesign wird Merkfähigkeit somit automatisch miterfasst, sollte aber auf Grund der kurzen Zeitspanne die Testleistung nur zu einem geringeren Ausmaß mitbestimmen als das logische Schlussfolgern. Die bisherigen Ergebnisse zum AcquA-Testdesign fallen diesbezüglich nicht eindeutig aus. So haben Hellwig et al. (2020) den Zusammenhang zwischen EU und verbaler Merkfähigkeit untersucht und eine mittlere latente Korrelation von .31 gefunden, wohingegen der latente Zusammenhang zum figuralen logischen Schlussfolgern mit .46 deskriptiv höher ausfiel. Schulze und Jobmann (2016) untersuchten den Zusammenhang zwischen EU und Merkfähigkeit sowie Arbeitsgedächtniskapazität, die eine hohe Überlappung mit dem logischen Schlussfolgern aufweist (Süß et al., 2002). Zur Erfassung von Merkfähigkeit wurden in der Studie verbale, numerische und figurale Aufgaben eingesetzt, die sich explizit auf das Aufgabenmaterial der AcquA-EU Aufgaben beziehen. Der latente Zusammenhang zwischen EU und Arbeitsgedächtniskapazität betrug .54 und der zwischen EU und Merkfähigkeit .64 (Schulze & Jobmann, 2016). Insgesamt lässt sich somit vermuten, dass Merkfähigkeit die Leistung bei den AcquA-Aufgaben mitbestimmt und daher auch bei den AcquA-PU Aufgaben untersucht werden sollte. Dabei ist zu beachten, dass logisches Schlussfolgern in den klassischen Inhaltsbereichen (verbal, numerisch und figural) und Merkfähigkeit ebenfalls hohe Zusammenhänge aufweisen (Süß et al., 2002). Allein aus diesem Grund ist ein mindestens mittlerer Zusammenhang der AcquA-PU Aufgaben mit einem Maß für Merkfähigkeit zu erwarten, sollte logisches Schlussfolgern tatsächlich die zentrale kognitive Anforderung der Aufgaben darstellen.

Das dritte Ziel ergab sich aus den bisherigen Ergebnissen zur Schwierigkeit der AcquA-PU Aufgaben. In den ersten beiden Studien zeigten ein Großteil der Items eine sehr geringe Schwierigkeit, was vor allem bei Verwendung des dichotomen Scorings deutlich wurde. Zwar zeigten in der ersten Studie auch einige AcquA-EU Aufgaben und Items sehr geringe

Schwierigkeiten (z.B. EU8 und EU9; vgl. Tabelle B3 in Anhang B) und auch in früheren Studien fielen einzelne Items durch solche Werte auf (Hellwig, 2016), allerdings nicht in einem so großen Ausmaß wie bei den AcquA-PU Aufgaben. Einerseits lässt sich dieses Ergebnis damit vereinbaren, dass in der Vergangenheit nur wenig interindividuelle Unterschiede in der akkuraten Beurteilung der Persönlichkeit anderer Personen gefunden werden konnten (Biesanz, 2010; Kenny, 1994; Funder, 1999) – andererseits ist nicht auszuschließen, dass dies ein Effekt der in den Studien beurteilten Zielpersonen war und diese nicht genügend relevante Hinweise auf ihre Persönlichkeitseigenschaften lieferten (Rogers & Biesanz, 2019). Letzteres kann bei den AcquA-PU Aufgaben durch deren Konstruktionsweise weitestgehend ausgeschlossen werden. Daher sollte in der dritten Studie ein erster Versuch einer theoretisch begründeten Manipulation der Aufgabenschwierigkeit umgesetzt werden. Auch Hellwig (2016) hat eine gezielte Konstruktion schwieriger Aufgaben vorgenommen, indem eine große Anzahl an Kontingenzen sowie ungewöhnliche Kontingenzen in die AcquA-EU Aufgaben integriert wurden. Hier zeigte sich insbesondere durch die Verwendung ungewöhnlicher Kontingenzen ein Effekt (siehe auch Hellwig et al., 2020). Wie bei Hellwig (2016) sollte eine Erhöhung der Schwierigkeit durch die gezielte Manipulation der Kontingenzen der Aneignungsphase erfolgen, da diese das Kernelement der AcquA-PU Aufgaben darstellen (Schulze & Roberts, 2015) und zur Modellierung der Persönlichkeit der Zielpersonen verwendet werden. Im Rahmen des AcquA-Testdesigns wurden verschiedene schwierigkeitsstiftende Elemente vorgeschlagen (Schulze & Roberts, 2015), von denen sich für PU aus theoretischer Sicht und auf Basis empirischer Evidenz aus relevanten Forschungsbereichen insbesondere zwei eignen: Die Verwendung von 1.) probabilistischen Kontingenzen sowie 2.) Kontingenzketten.

Bei probabilistischen Kontingenzen zeigt die Zielperson mit einer relativen Häufigkeit $0 < p < 1$ eine bestimmte Reaktion auf das auslösende Ereignis (Schulze & Roberts, 2015), sodass hierdurch inkonsistentes Verhalten der Zielperson modelliert werden kann. Induktives logisches Schlussfolgern erfordert das Verstehen der Prämissen und die Bildung einer mentalen Repräsentation der vorliegenden Problemstellung (Johnson-Laird, 1994; Primi, 2001) beziehungsweise das Erkennen und Verstehen der Regelhaftigkeiten (Schneider & McGrew, 2018). Dies entspricht bei PU dem Erkennen und Verstehen der typischen Reaktionen auf bestimmte Ereignisse. Sind diese Regelhaftigkeiten probabilistisch anstatt deterministisch, sollte das Erkennen und Verstehen erschwert und als Folge die Aufgabenschwierigkeiten erhöht sein. Entsprechende Überlegungen finden sich auch im Bereich akkurater Persönlichkeitsbeurteilungen. Wie Human und Biesanz (2013) in einem integrativen Review

über die gute Zielperson zusammengefasst haben, kann angenommen werden, dass eine über verschiedene Situationen hinweg konsistente und zeitlich stabile Persönlichkeit sowie entsprechendes Verhalten akkuratere Beurteilungen erleichtert, da Zielpersonen mit einer solch kohärenten Persönlichkeit mehr relevante Verhaltensweisen zeigen. Diese Annahme stützten Human und Biesanz (2013) zunächst durch eine Reihe an empirischen Befunden anderer Autor:innen. Später führten die Autorin und der Autor zusammen mit weiteren Kolleg:innen zudem eigene Studien durch (Human et al., 2014; Human et al., 2019), die ebenfalls darauf hindeuten, dass Zielpersonen mit einer kohärenten Persönlichkeit mit einer höheren Akkuratheit beurteilt werden. Nach Human und Biesanz (2013; vgl. auch Human et al., 2014; Human et al., 2019) ist dieser Aspekt dem ersten Schritt im RAM zuzuordnen, in dem die entscheidende Frage ist, ob die Zielperson etwas für die Persönlichkeitseigenschaft Relevantes tut (Funder, 1995). Demnach könnte man argumentieren, dass Zielpersonen mit probabilistischen Kontingenzen schlechte Zielpersonen darstellen, bei denen sich entsprechend den Ergebnissen von Rogers und Biesanz (2019) möglicherweise keine interindividuellen Fähigkeitsunterscheide zeigen. Nach Rogers und Biesanz (2019) und in Übereinstimmung mit dem RAM ist allerdings der entscheidende Punkt, ob die Zielperson auch wirklich relevante Hinweise zeigt oder nicht. Probabilistische Kontingenzen liefern solche relevanten Hinweise, wenngleich mit dem Unterschied, dass die Reaktion der Zielperson nur mit einer bestimmten Wahrscheinlichkeit gezeigt wird. Unter dieser Annahme ließe sich dieser Aspekt auch dem letzten RAM-Schritt (Nutzung) zuordnen, in dem Fähigkeitsunterschiede der Beurteiler:innen eine entscheidende Rolle spielen (Funder, 1995). Hier muss die wahrgenommene Information unter anderem richtig miteinander kombiniert werden (Letzring & Funder, 2021), wozu das Erkennen der probabilistischen Natur des relevanten Verhaltens gezählt werden kann. Insgesamt betrachtet wurde durch die Verwendung probabilistischer Kontingenzen eine Steigerung der Aufgabenschwierigkeit erwartet.

Bei Kontingenzketten fungiert die Reaktion aus einer ersten Kontingenz als Ereignis für eine weitere Kontingenz, wobei die Länge solcher Ketten variieren kann (Schulze & Roberts, 2015). Über die Kettenlänge solcher Ereignis-Reaktions-Beziehungen lässt sich die Komplexität des Verhaltens der Zielperson modellieren und zwar vor allem dann, wenn verschiedene Persönlichkeitseigenschaften die Grundlage dieser Reaktionen darstellen. Dass eine Erhöhung der Kettenlänge auch die Aufgabenschwierigkeit erhöhen sollte, lässt sich über die angenommene zentrale kognitive Operation von PU (logisches Schlussfolgern) und deren Operationalisierung in klassischen Inhaltsbereichen herleiten. Wie bei den probabilistischen Kontingenzen sollte auch die Verwendung von Kontingenzketten dazu führen, dass das

Erkennen und Verstehen der Prämissen (Johnson-Laird, 1994) beziehungsweise der Regelhaftigkeiten (Schneider & McGrew, 2018) erschwert ist, da sich die typische Reaktion der Zielperson aus mehreren Teilreaktionen zusammensetzt. Bei Aufgaben zur Erfassung von figuralem und numerischem logischen Schlussfolgern (z.B. Matrizen und Zahlenreihen) wurden die Anzahl der bei einem Item zu verarbeitenden Elemente und die Anzahl beziehungsweise Komplexität der Beziehungen unter den Elementen als schwierigkeitsstiftende Eigenschaften angenommen und untersucht (Freund et al., 2008; Holzman et al., 1983; Loe et al., 2018; Primi, 2001). Die Annahme der zitierten Studien war, dass bei einer höheren Anzahl an Elementen und Beziehungen mehr Information verarbeitet werden muss und damit auch eine höhere kognitive Kapazität erforderlich ist beziehungsweise das Arbeitsgedächtnis stärker beansprucht wird. Insgesamt weisen die Studienergebnisse darauf hin, dass die zu verarbeitende Informationsmenge die Itemschwierigkeit erhöht (Freund et al., 2008; Primi, 2001; Holzman et al., 1983; Loe et al., 2018), während die alleinige Betrachtung der Anzahl der Elemente und Anzahl der Regeln nicht überall einen signifikanten Einfluss auf die Schwierigkeit hatte (Primi, 2001). Durch Verwendung von Kontingenzketten lässt sich bei den AcquA-PU Aufgaben ebenfalls die Informationsmenge steigern, da sowohl die Anzahl der zu verarbeitenden Elemente (Ereignisse und Reaktionen) als auch die Anzahl der Beziehungen zwischen den Elementen (Kontingenzen) erhöht wird. Aus diesem Grund wurde durch die Verwendung von Kontingenzketten ebenfalls eine Steigerung der Aufgabenschwierigkeit erwartet.

6.2 Methode

6.2.1 Stichprobe

Insgesamt haben 171 Personen an der Studie teilgenommen, von denen allerdings neun aus den folgenden Gründen ausgeschlossen werden mussten: 1.) Drei Personen haben an der Vorgängerstudie teilgenommen und wurden auf Grund der großen inhaltlichen Überlappung der beiden Studien und den damit verbundenen erwartbaren Erinnerungs- und Übungseffekten ausgeschlossen.[14] 2) Vier Personen mussten auf Grund des Sprachkriteriums (mindestens 10 Jahre Deutschkenntnisse; vgl. Vorgängerstudien) ausgeschlossen werden, da die Studie erneut

[14] Sieben Personen, die an der ersten Studie der Studienreihe teilgenommen haben, wurden hingegen nicht ausgeschlossen. Die inhaltliche Überlappung der beiden Studien war auf Grund von Aufgabenüberarbeitungen und Neukonstruktionen deutlich geringer und der zeitliche Abstand mit mindestens 1.5 Jahren zudem groß genug.

einen hohen sprachlichen Anteil aufwies. 3) Zwei Personen haben bei mindestens einem der eingesetzten Instrumente die Aufgabeninstruktion missachtet.

Die Analysestichprobe bestand somit aus N = 162 Personen (112 Frauen, 50 Männer), die über Aushänge an der Bergischen Universität Wuppertal und Social Media rekrutiert wurden. Zum ersten Erhebungszeitpunkt lag das mittlere Alter bei 24.39 Jahren (SD = 7.47; Range: 18 bis 67). Deutsch als Muttersprache gaben 138 Personen an und diejenigen Personen, die eine andere Muttersprache nannten, sprachen seit mindestens 11 Jahren Deutsch. Bei der Frage nach dem höchsten akademischen oder schulischen Abschluss wurden folgende Antwortoptionen ausgewählt: Fachhochschul- oder Hochschulreife/Abitur (n = 126), Bachelor (n = 25), Master/ Diplom (n = 8), Realschul- oder gleichwertiger Abschluss (n = 3). Der weitaus größte Teil der Analysestichprobe bestand aus Studierenden unterschiedlicher Fachrichtungen (n = 149), wobei die Psychologiestudierenden mit 65 Personen am stärksten vertreten waren.

Da die Studie erst nachträglich um das Instrument zur Erfassung von logischem Schlussfolgern und Merkfähigkeit ergänzt wurde und darüber hinaus bei zwei Personen nur einer der zwei Erhebungszeitpunkte stattfinden konnte, haben nicht alle Personen alle Instrumente bearbeitet. Vollständige Daten zu allen Instrumenten lagen nur für eine Teilstichprobe von 129 Personen vor, die aus 95 Frauen und 34 Männern bestand. Zum ersten Erhebungszeitpunkt lag das mittlere Alter dieser Teilstichprobe bei 23.91 Jahren (SD = 6.39; Range: 18 bis 67). Die Muttersprache Deutsch gaben 108 Personen an und die übrigen Personen sprachen seit mindestens 11 Jahren Deutsch. Zudem verteilten sich die Angaben zum höchsten akademischen oder schulischen Abschluss wie folgt: Fachhochschul- oder Hochschulreife/Abitur (n = 102), Bachelor (n = 18), Master/ Diplom (n = 6), Realschul- oder gleichwertiger Abschluss (n = 3). Die Teilstichprobe bestand aus 121 Studierenden, von denen 61 das Studienfach Psychologie nannten.

6.2.2 Messinstrumente

6.2.2.1 Personality Understanding. Zur Erfassung von PU wurden insgesamt fünf AcquA-PU Aufgaben eingesetzt (vgl. Tabelle 6.1). Drei dieser Aufgaben wurden unverändert aus Studie 2 übernommen. Hierbei handelte es sich um die Aufgaben PU2, PU3 und PU6. Die Aufgaben PU1 und PU5 wurden in einer überarbeiteten Form eingesetzt. Entsprechend des dritten Ziels der Studie wurde eine gezielte Manipulation der Schwierigkeit vorgenommen, indem probabilistische Kontingenzen (PU1) beziehungsweise Kontingenzketten (PU5) integriert wurden. Zur Bestimmung der Itemscores der AcquA-PU Aufgaben wurde erneut

primär das standardisierte Distanzscoring verwendet, dessen Ergebnisse durch das dichotome Scoring ergänzt wurden.

Tabelle 6.1

Übersicht über verschiedene Aspekte der in Studie 3 eingesetzten AcquA-Aufgaben zur Erfassung von Personality Understanding (PU)

		Anzahl		
Aufgabe[a]	AB5C-Facette der Kontingenzen	Zielpersonen in AP1	relevante Kontingenzen in AP1	Einschätzungen in AP2
PU1p	Efficiency	2	2	12
PU2	Orderliness	2	2	12
PU3	Conscientiousness	2	4	24
PU5k	Pleasantness	2	2	10
PU6	Cooperation	2	2	12
Gesamt		10	12	70

Anmerkungen. AP1 = Aneignungsphase; AP2 = Anwendungsphase.

[a] Alle Aufgaben wurden in der Version 2 eingesetzt, d.h. die Dialoge wurden ohne den Ausdruck einer bestimmten Emotion vertont.

Konstruktion probabilistischer Kontingenzen. Zur Konstruktion der probabilistischen Kontingenzen sollte Aufgabe PU1 so überarbeitet werden, dass das auslösende Ereignis der relevanten Kontingenz im Laufe der Aneignungsphase mehrfach präsentiert wird. Auf das Ereignis sollten zudem zwei verschiedene Reaktionen folgen: Eine wahrscheinliche und eine unwahrscheinliche Reaktion. Die wahrscheinliche Reaktion folgt auf das Ereignis mit einer relativen Häufigkeit $p > .5$, die unwahrscheinliche Reaktion mit der relativen Häufigkeit $1 – p$.

Diese Art der Umsetzung probabilistischer Kontingenzen hätte je nach gewählter relativer Häufigkeit der Reaktionen eine große Anzahl an Kontingenzpräsentationen erfordert, was eine deutliche Verlängerung der Aneignungsphase und somit eine höhere Beanspruchung der Merkfähigkeit der Testpersonen zur Folge gehabt hätte. Zudem war bei zu langen Aufgaben ein Motivationsverlust der Testpersonen zu erwarten. Aus diesem Grund wurden bei Aufgabe PU1 nur drei Kontingenzpräsentationen umgesetzt. Bei zwei der drei Präsentationen zeigt die Zielperson die wahrscheinliche Reaktion A (relative Häufigkeit: $.\overline{6}$) und in einer der drei Situationen die unwahrscheinliche Reaktion B (relative Häufigkeit: $.\overline{3}$). In der

Anwendungsphase müssen von den Testpersonen schließlich Reaktionen eingeschätzt werden, die entweder Reaktion A entsprechen (korrekte Antwort auf der bei den AcquA-PU Aufgaben verwendeten Skala von 0 bis 10: $6.\overline{6}$) oder Reaktion B (korrekte Antwort: $3.\overline{3}$) oder aber weder Reaktion A noch Reaktion B (korrekte Antwort: 0).

Abgesehen von der Umsetzung der probabilistischen Kontingenzen wurde darauf geachtet, keine weiteren inhaltlichen Änderungen vorzunehmen beziehungsweise nur dann, wenn es für die Modellierung der Kontingenzen zwingend erforderlich war. Dies ermöglichte einen deskriptiven Vergleich mit den Ergebnissen aus Studie 2 und somit eine erste Überprüfung, ob probabilistische Kontingenzen die Aufgabenschwierigkeit erhöhen können. Als neuer Name für die Aufgabe wurde die Bezeichnung PU1p gewählt.

Konstruktion von Kontingenzketten. Bei Aufgabe PU5 wurden im Rahmen der Überarbeitung durch die Verwendung von Kontingenzketten komplexe relevante Kontingenzen konstruiert (neuer Name der Aufgabe: PU5k). Diese bestehen aus mehreren Teilreaktionen und können inhaltlich mehreren Persönlichkeitseigenschaften zugeordnet werden. Während die relevanten Kontingenzen in Studie 2 noch ausschließlich Verträglichkeit widerspiegelten, wurden für die Kontingenzketten der vorliegenden Studie zusätzlich Verhaltensweisen der Extraversion beziehungsweise Gewissenhaftigkeit integriert.

Die erste Kontingenzkette startet mit einem auslösenden Ereignis für extravertiertes Verhalten (niedrige Ausprägung) der Zielperson. Dieses initiiert wiederum die Reaktion einer anderen Person, die als auslösendes Ereignis für das verträgliche Verhalten (hohe Ausprägung) agiert, das mit einem Abbruch des extravertierten Verhaltes einhergeht. Die Komplexität besteht bei dieser Kontingenz somit darin, dass das extravertierte Verhalten durch eine verträgliche Verhaltensweise abgelöst wird. Zudem ist an dieser Kontingenzkette das Verhalten einer weiteren Person beteiligt. Die zweite Kontingenzkette kommt ohne weitere Person aus, ähnelt aber ansonsten der ersten Kette. Start ist das auslösende Ereignis für verträgliches Verhalten (niedrige Ausprägung), was letztendlich in dem auslösenden Ereignis für gewissenhaftes Verhalten (hohe Ausprägung) resultiert und dieses initiiert.

Die Umsetzung der Kontingenzketten erforderte eine umfangreichere Überarbeitung der Aufgabe PU5, was in Anbetracht der unzureichenden psychometrischen Ergebnisse aus Studie 2 auch angemessen erschien und weswegen Aufgabe PU5 hierfür ausgewählt wurde. Ein Vergleich mit den Ergebnissen der Studie 2 war daher nicht sinnvoll.

6.2.2.2 Social Understanding. Zur Erfassung von SU wurden drei Aufgaben aus der dritten Version des MTSI (MTSI-3; Süß et al., 2009) eingesetzt. Der MTSI-3 ist ein computerbasierter Leistungstest, der auf Basis des Integrativen Modells Sozialer Intelligenz von Weis und Süß (2005; Weis et al., 2006) entwickelt wurde und Aufgaben zur Erfassung mehrerer Teilfähigkeiten der SI beinhaltet. Der Aufbau der SU-Aufgaben wird im Folgenden auf Basis von Süß et al. (2009; vgl. auch Baumgarten, 2015; Conzelmann et al., 2013) beschrieben: Im Fokus jeder Aufgabe steht eine reale Zielperson. Über diese erhalten die Testpersonen zu Beginn der Aufgabe zunächst allgemeine Information. Im Laufe der Aufgabe wird die Zielperson in verschiedenen sozialen Situationen präsentiert, wobei die Präsentation der Information entweder verbal (Text), auditiv (Ton), als statische Bilder (Fotos) oder als bewegte Bilder (Videos ohne Ton) erfolgt. Im Anschluss an jede soziale Situation werden die Testpersonen gebeten, die Zielperson im Hinblick auf ihre Emotionen, Kognitionen und Beziehungen zu anderen Personen auf einer 7-stufigen Likert-Skala einzuschätzen. Am Ende jeder Aufgabe erfolgen zudem einige globale Einschätzungen der Zielperson, wie zum Beispiel ihrer Persönlichkeit auf Basis der Big Five-Faktoren, die allerdings nicht in die standardmäßige Auswertung der SU-Aufgaben einfließen. Die für die Auswertung der SU-Aufgaben benötigten richtigen Antworten wurden beim MTSI-3 auf Basis des Target-Scorings generiert, stellen also die Selbsteinschätzungen der Zielpersonen dar.

Die vorhandene Validitätsevidenz der SU-Aufgaben des MTSI-3 und seiner Vorgängerversion MTSI-2 ist relativ begrenzt und bezieht sich vor allem auf die faktorielle Validität sowie auf Zusammenhänge zu anderen Teilfähigkeiten der SI, Maßen der klassischen Intelligenz (logisches Schlussfolgern, Merkfähigkeit, Bearbeitungsgeschwindigkeit) sowie zur Persönlichkeit der Testpersonen. Es zeigten sich Anhaltspunkte für Eindimensionalität der SU-Aufgaben, maximal geringe Zusammenhänge zum Sozialen Gedächtnis, hohe Zusammenhänge zum Sozialen Wissen, maximal geringe Zusammenhänge zu Maßen der klassischen Intelligenz sowie bei einigen SU-Aufgaben kleine bis mittlere Zusammenhänge zu Extraversion, Gewissenhaftigkeit und Verträglichkeit (Baumgarten, 2015; Conzelmann et al., 2013; Seidel, 2007; Weis, 2008). Die Validität der SU-Aufgaben kann also noch nicht abschließend beurteilt werden, worauf auch die Autor:innen der bisherigen Studien hinweisen (z.B. Baumgarten, 2015; Conzelmann et al., 2013). Darüber hinaus ist das verwendete Target-Scoring problembehaftet und stellt kein eindeutiges, rationales Scoring der Testpersonenantworten dar (vgl. Abschnitt 3.2). Wie die Autor:innen des MTSI anmerken, kann beispielsweise nicht ausgeschlossen werden, dass die Zielpersonen sozial erwünscht geantwortet haben oder nicht fähig waren, korrekte Antworten zu geben (Conzelmann et al.,

2013). Trotz dieser Einschränkungen wurde der MTSI-3 zur Operationalisierung von SU ausgewählt, da kein alternatives und besseres Verfahren aus dem Bereich der Leistungstests zur Verfügung stand. Zudem weist der MTSI-3 eine eindeutige theoretische Grundlage auf und die Konstruktion der SU-Aufgaben wurde dem Konstrukt angemessen vorgenommen. Insbesondere das realistische Testmaterial, das erstellt wurde, indem echte Personen in ihrem normalen Alltag mit der Videokamera begleitet wurden (Conzelmann et al., 2013), ist hier positiv hervorzuheben.

Die Auswahl der SU-Aufgaben erfolgte auf Basis bisheriger Ergebnisse zur Reliabilität der Aufgaben sowie deren Ladungen im Faktormodell (Baumgarten, 2015; Conzelmann et al., 2013). Darüber hinaus standen zum Zeitpunkt der Aufgabenauswahl nicht mehr alle Aufgaben zur Verfügung (M. Baumgarten, persönliche Kommunikation, 12. September, 2018). Ausgewählt wurden schließlich die drei Zielpersonen RF, FB sowie HR, die zusammen mit jeweils einer spezifischen und einer allgemeinen Instruktion eingesetzt wurden. Sämtliche Instruktionen zu den SU-Aufgaben wurden vorab vertont und den Testpersonen standardisiert und computerbasiert mit Hilfe eines separaten Laptops schriftlich und zusätzlich auditiv dargeboten. In Tabelle 6.2 sind demografische Daten der Zielpersonen sowie Reliabilitätsschätzungen aus bisherigen Studien zusammengefasst. Die geschätzte Durchführungsdauer je Zielperson liegt bei 20 bis 25 Minuten (Süß et al., 2009).

Tabelle 6.2

Demografische Daten der Zielpersonen des MTSI-3 aus Weis (2008) und Baumgarten (2015) sowie Schätzungen der Reliabilität (Cronbachs α) aus bisherigen Studien

Zielperson	Geschlecht	Alter	Beruf	α
RF	weiblich	24	medizinisch-technische Laborassistentin	.74[a]
FB	männlich	69	Immobilienkaufmann	.78[a]
				.50[b]
HR	weiblich	60	Lehrerin	.86[a]
				.69[b]

Anmerkungen. [a] Aus Conzelmann et al. (2013; Studie 2). Die Zuordnung der Ergebnisse aus der Studie zu den Zielpersonen wurde durch Nachfrage bei S. Weis vorgenommen (M. Baumgarten, persönliche Kommunikation, 28. August 2018). Die Reliabilitätsschätzung erfolgte auf Ebene der einzelnen Itemscores (reduzierter Itempool). [b] Aus Baumgarten (2015, Studie 1). Schätzung erfolgte auf Ebene aller Situationsscores.

6.2.2.3 Selbst-Fremd Übereinstimmung. Die Bestimmung der Akkuratheit von Persönlichkeitsbeurteilungen mittels Selbst-Fremd Übereinstimmung erfolgte über eine Erweiterung der SU-Aufgaben des MTSI-3. Dies bot sich an, da im Rahmen der Aufgabenkonstruktion der SU-Aufgaben Selbsteinschätzungen der Zielpersonen zu den Big Five-Faktoren mit Hilfe der deutschsprachigen Version des NEO-FFI von Borkenau und Ostendorf (2008) erhoben und für die vorliegende Studie zur Verfügung gestellt wurden (Süß et al., 2009; Weis, 2008). Bei den SU-Aufgaben sind globale Persönlichkeitsbeurteilungen der Zielpersonen mit Hilfe eines Items je Big Five-Faktor vorgesehen (Süß et al., 2009). Diese 1-Item-Erhebungen wurden für das Studienziel als nicht reliabel und valide genug erachtet. Daher wurde im Anschluss an jede SU-Aufgabe eine Fremdbeurteilung der Persönlichkeit der Zielperson mit Hilfe des NEO-FFI erhoben. Da im NEO-FFI eigentlich keine Fremdberichtsform vorgesehen ist (vgl. Borkenau & Ostendorf, 2008), wurde diese analog zur Fremdberichtsform des NEO-PI-R (Ostendorf & Angleitner, 2004) erstellt, der die Items des NEO-FFI vollständig beinhaltet (Borkenau & Ostendorf, 2008).

6.2.2.4 Persönlichkeit der Testpersonen. Der NEO-FFI von Borkenau und Ostendorf (2008) wurde zudem in der Selbstberichtsform zur Erfassung der Persönlichkeit der Testpersonen eingesetzt. Dies erfolgte nicht im Rahmen eines der Studienziele, sondern auf Grund anderer Aspekte der Studiendurchführung.

6.2.2.5 Logisches Schlussfolgern. Zur Erfassung von logischem Schlussfolgern wurde wie in Studie 2 das Modul 4 aus dem WIT-2 von Kersting et al. (2008) eingesetzt. Die Fähigkeit zum logischen Schlussfolgern wird hier unter der Bezeichnung schlussfolgerndes Denken durch die drei Untertests Analogien, Abwicklungen und Zahlenreihen erfasst.

6.2.2.6 Merkfähigkeit. Zur Erfassung der Merkfähigkeit wurde Modul 5 aus dem WIT-2 (Kersting et al., 2008) eingesetzt. Das Modul umfasst eine Einprägphase (4 Minuten) und eine Reproduktionsphase (3 Minuten), die von anderen Untertests für etwa 14 bis 20 Minuten unterbrochen werden. Das einzuprägende Material besteht aus verbaler, numerischer sowie figuraler Information und wird in der Reproduktionsphase mit Hilfe von 21 Multiple-Choice-Items abgefragt. Die Reliabilität liegt laut Manual bei Cronbachs $\alpha = .78$ (Kersting et

al., 2008). In der vorliegenden Studie wurde als Störaufgabe der Untertest Abwicklungen aus Modul 4 verwendet, für den laut Manual etwa 14 Minuten vorgesehen sind.

6.2.3 Durchführung

Die Datenerhebungen, die auf Grund der COVID-19-Pandemie mehrfach und für längere Zeiträume unterbrochen werden mussten, wurden in ruhigen Laborräumen der Bergischen Universität Wuppertal durchgeführt. An einem Erhebungstermin konnten grundsätzlich bis zu vier Testpersonen gleichzeitig teilnehmen, allerdings wurden auf Grund der Pandemiesituation hauptsächlich Einzelerhebungen durchgeführt.

Während die Aufgaben des WIT-2 und die Selbstberichtsversion des NEO-FFI im Paper-Pencil-Format durchgeführt wurden, wurden die AcquA-PU Aufgaben, die SU-Aufgaben des MTSI-3 sowie der NEO-FFI in der Fremdberichtsversion computerbasiert präsentiert. Gleiches gilt für die Erhebung der Demografie und die Instruktionen zu den SU-Aufgaben. Die Präsentation der SU-Aufgaben erfolgte mit einer eigens hierfür geschriebenen Software (Süß et al., 2009), für die übrige Programmierung und Darbietung wurde Inquisit 4 (Millisecond Software, 2016) verwendet. Die auditive Präsentation der Instruktionen zu den AcquA-PU Aufgaben und SU-Aufgaben erfolgte über einen Kopfhörer, sodass individuelle Bearbeitungsgeschwindigkeiten realisiert werden konnten.[15]

Jeder Erhebungstermin startete mit einer schriftlichen allgemeinen Information zur Studie sowie einer schriftlichen Einwilligung der Testpersonen in die Studienteilnahme. Die Studie begann anschließend durch das Vorlesen der allgemeinen Instruktion durch die Versuchsleitung, die zusätzlich schriftlich am Computer präsentiert wurde. Im Anschluss wurden die Testpersonen gebeten, am Computer einige demografische Angaben zu machen. Bei der Darstellung der restlichen Durchführung müssen im Folgenden zwei Versionen unterschieden werden, da die Studie erst nachträglich, das heißt nach der Erhebung von etwa 50 Testpersonen, um die Aufgaben des WIT-2 ergänzt wurde.

[15] Bei den Einzelerhebungen unter Pandemiebedingungen wurde auf den Kopfhörer verzichtet. Hier erfolgte die auditive Präsentation über die Lautsprecher des Computers oder Laptops.

6.2.3.1 Reihenfolge Version 1. Der Ablauf der Untersuchung sowie Art der Administration und Zeitschätzungen der einzelnen Instrumente vor Hinzunahme der Aufgaben des WIT-2 kann Tabelle 6.3 entnommen werden.

Tabelle 6.3

Reihenfolge, Art der Administration und Zeitschätzung (in Minuten) der in Studie 3 eingesetzten Instrumente (Version 1; ohne Aufgaben des WIT-2)

Instrument	Administration	Zeitschätzung
1. Allgemeine Instruktion & Demografie	Computer	7
2. NEO-FFI (Selbstbericht)	Paper-Pencil	7
3. AcquA-PU Teil 1 (Instruktion, 3 Aufgaben)	Computer	40
Pause (ca. 5 Minuten; optional)		
4. AcquA-PU Teil 2 (2 Aufgaben)	Computer	20
5. MTSI-SU, Allgemeine Instruktion	Laptop	5
6. MTSI-SU, Zielperson HR	Computer, Laptop	25
7. NEO-FFI, Fremdbericht HR	Laptop	7
Pause (ca. 5 Minuten; optional)		
8. MTSI-SU, Zielperson FB	Computer, Laptop	25
9. NEO-FFI, Fremdbericht FB	Laptop	7
10. MTSI-SU, Zielperson RF	Computer, Laptop	25
11. NEO-FFI, Fremdbericht RF	Laptop	7

Anmerkung. Die geschätzte Studiendauer für Version 1 betrug insgesamt 185 Minuten (inkl. Pausen).

Die AcquA-PU Aufgaben wurden in einer unvollständig ausbalancierten Reihenfolge präsentiert. Hierfür wurden fünf mögliche Reihenfolgen konstruiert, bei denen darauf geachtet wurde, dass die beiden überarbeiteten Aufgaben PU1p und PU5k nicht direkt hintereinander präsentiert werden. Zu diesen fünf Reihenfolgen wurden die Testpersonen zufällig zugewiesen.

Wie bereits geschildert, wurden die Instruktionen der SU-Aufgaben den Testpersonen sowohl schriftlich als auch auditiv präsentiert. Dieses Vorgehen wurde gewählt, um bei Gruppenerhebungen individuelle Bearbeitungsgeschwindigkeiten zu ermöglichen und Wartezeiten zu vermeiden. Da die SU-Aufgaben allerdings mit einer eigenen

Computersoftware präsentiert werden mussten, musste die Präsentation der Instruktionen auf einem separaten Laptop erfolgen. Auch die Erfassung des NEO-FFI Fremdberichts der Zielpersonen erfolgte aus Gründen besserer Umsetzbarkeit an dem separaten Laptop. Die Testpersonen mussten somit im Laufe der Erhebung mehrfach zwischen einem Computer mit externem Monitor und einem danebenstehenden Laptop hin und her wechseln. Die Zeitpunkte für die Wechsel wurden ausführlich in der Instruktion erklärt und der Ablauf durch die anwesende Versuchsleitung überwacht.

Nach Abschluss der Erhebung konnten die Testpersonen zwischen drei möglichen Aufwandsentschädigungen wählen: 1.) Erhalt von 20 € in bar; 2.) Rückmeldung zur eigenen Persönlichkeit auf Basis des NEO-FFI Selbstberichts; 3.) Gutschrift von Versuchspersonenstunden entsprechend der tatsächlichen Studiendauer. Insgesamt variierte die Studiendauer bei Version 1 zwischen 2:15 h und 3:30 h.

Nach Erweiterung der Studie um die Aufgaben aus dem WIT-2 konnten 13 der bereits erhobenen Testpersonen für eine Nacherhebung des schlussfolgernden Denkens und der Merkfähigkeit rekrutiert werden. Diese Testpersonen erhielten für die Nacherhebung 10 € in bar. Darüber hinaus fand kurz vor der hier beschriebenen Studie eine weitere Untersuchung am Lehrstuhl für Methodenlehre und Psychologische Diagnostik der Bergischen Universität Wuppertal statt, in der dieselben Aufgaben des WIT-2 eingesetzt wurden.[16] Für zwei weitere Testpersonen, die an beiden Untersuchungen teilnahmen, konnten die Daten der WIT-2 Aufgaben aus dieser weiteren Untersuchung verwendet werden.

6.2.3.2 Reihenfolge Version 2. Durch Hinzunahme der Aufgaben aus dem WIT-2 verlängerte sich die geschätzte Studiendauer um fast 50 Minuten, sodass die Studie auf zwei Erhebungstermine aufgeteilt wurde. Eine Übersicht über den Ablauf der Untersuchung sowie Art der Administration und Zeitschätzungen der eingesetzten Instrumente in dieser zweiten Version der Studie sind Tabelle 6.4 zu entnehmen. Hierbei ist anzumerken, dass im Falle einer Gruppenerhebung nur die Aufgaben des WIT-2 auf Grund der einzuhaltenden Bearbeitungszeiten in der Gruppe instruiert wurden. Die übrigen Aufgaben bearbeiteten die Testpersonen in individueller Geschwindigkeit.

[16] Bei der Untersuchung handelte es sich um die Studie zur Masterarbeit von Shabnam Resasade.

Tabelle 6.4

Reihenfolge, Art der Administration und Zeitschätzung (in Minuten) der in Studie 3 eingesetzten Instrumente (Version 2; mit Aufgaben des WIT-2)

Termin/Instrument	Administration	Zeitschätzung
Termin 1		
1. Allgemeine Instruktion & Demografie	Computer	7
2. NEO-FFI (Selbstbericht)	Paper-Pencil	7
3. MTSI-SU, Allgemeine Instruktion	Laptop	5
4. MTSI-SU, Zielperson HR	Computer, Laptop	25
5. NEO-FFI, Fremdbericht HR	Laptop	7
Pause (ca. 5 Minuten; optional)		
6. AcquA-PU (Instruktion, 5 Aufgaben)	Computer	60
Termin 2		
1. Allgemeine Instruktion & Demografie	Computer	5
2. WIT-2, Allgemeine Instruktion	Paper-Pencil	5
3. WIT-2, Analogien und Zahlenreihen	Paper-Pencil	20.5
4. WIT-2, Merkfähigkeit (Einprägphase)	Paper-Pencil	4.5
5. WIT-2, Abwicklungen	Paper-Pencil	14
6. WIT-2, Merkfähigkeit (Abrufphase)	Paper-Pencil	4.5
Pause (ca. 5 Minuten; optional)		
7. MTSI-SU, Allgemeine Instruktion	Laptop	5
8. MTSI-SU, Zielperson FB	Computer, Laptop	25
9. NEO-FFI, Fremdbericht FB	Laptop	7
10. MTSI-SU, Zielperson RF	Computer, Laptop	25
11. NEO-FFI, Fremdbericht RF	Laptop	7

Anmerkungen. Die geschätzte Studiendauer für Termin 1 betrug insgesamt 116 Minuten (inkl. Pause) und für Termin 2 insgesamt 127.5 Minuten (inkl. Pause). Die Zeitschätzungen für die Aufgaben des WIT-2 wurden aus dem Manual (Kersting et al., 2008) übernommen.

Abgesehen von der Hinzunahme der WIT-2 Aufgaben und der Aufteilung auf zwei Erhebungstermine gab es keine Veränderungen. Die Testpersonen, die an Version 2 teilnahmen, konnten nach Abschluss beider Termine zwischen drei möglichen Aufwandsentschädigungen wählen: 1.) Erhalt von 40 € in bar; 2.) Rückmeldung zur eigenen Persönlichkeit auf Basis des NEO-FFI Selbstberichts; 3.) Versuchspersonenstunden entsprechend der tatsächlichen Studiendauer. Insgesamt variierte die Studiendauer der einzelnen Termine beider Erhebungszeitpunkte etwa zwischen 1:30 h und 2:30 h.

Wie bereits geschildert, fand am Lehrstuhl für Methodenlehre und Psychologische Diagnostik kurz vor der hier beschriebenen Studie eine weitere Erhebung statt, in der dieselben Aufgaben des WIT-2 eingesetzt wurden (vgl. Fußnote 16). Neun Testpersonen aus Studie 3 (Version 2) nahmen an beiden Studien teil und bearbeiteten somit die Aufgaben des WIT-2 zweimal hintereinander mit einem Abstand von < 1 bis 4.5 Monaten. Da von Übungs- und Erinnerungseffekten ausgegangen werden musste, wurden für diese neun Testpersonen die WIT-2 Daten der früheren Untersuchung verwendet.

6.2.4 Statistische Analysen

6.2.4.1 Allgemeines. Die Überprüfung der angenommenen Zusammenhänge von PU und SU, der Selbst-Fremd Übereinstimmung, schlussfolgerndem Denken sowie Merkfähigkeit erfolgte auf latenter Ebene mittels CFAs. Im Messmodell für PU wurden, wie in den vorangegangenen Studien, auf Grund der fehlenden lokalen stochastischen Unabhängigkeit der Items Aufgaben-Parcels als Indikatoren verwendet (vgl. Hellwig et al., 2020). Die latenten Zusammenhänge wurden auf Grund der relativ kleinen und zudem variierenden Größe der zur Verfügung stehenden Analysestichprobe in drei getrennten Modellen geschätzt. Im ersten Modell wurde der Zusammenhang zwischen PU und SU geschätzt, im zweiten der Zusammenhang zwischen PU und der Selbst-Fremd Übereinstimmung und im dritten Modell der Zusammenhang zwischen PU, schlussfolgerndem Denken sowie Merkfähigkeit. Für die Schätzung der CFA-Modelle wurde sowohl Mplus (Muthén & Muthén, 2017) als auch das R-Paket lavaan (Rosseel, 2012) verwendet. Die Reliabilität wurde mit Hilfe des R-Pakets MBESS (Kelley, 2007, 2020) geschätzt. Als Konfidenzintervalle für Cronbachs α sowie McDonalds ω wurden – falls nicht anders angegeben – Bootstrap-Perzentil-Intervalle mit 10000 Replikationen verwendet (vgl. Kelley, 2020; Kelley & Pornprasertmanit, 2016). Die hierfür notwendige Eindimensionalität wurde bei den Konstrukten, die mit bereits vorhandenen, standardisierten Testverfahren erfasst wurden (SU, schlussfolgerndes Denken, Merkfähigkeit),

angenommen. Bei PU sowie der Selbst-Fremd Übereinstimmung wurde die Eindimensionalität zuvor überprüft.

6.2.4.2 Zusammenhang PU und SU. Zur Untersuchung des Zusammenhangs zwischen PU und SU wurde ein zweifaktorielles CFA-Modell spezifiziert. Bei der Wahl der Indikatoren für SU musste beachtet werden, dass bei den SU-Aufgaben des MTSI-3 eine analoge Problematik bestand wie bei den AcquA-PU Aufgaben. Im Zentrum einer SU-Aufgabe steht stets eine Zielperson, die in mehreren Situationen präsentiert wird. Im Anschluss an jede Situation müssen in der Regel mehrere Items von den Testpersonen beantwortet werden, die sich alle auf die vorangegangene Situation beziehen (Süß et al., 2009). Es ist also von mehreren Abhängigkeiten der Items innerhalb einer SU-Aufgabe auszugehen, die von Baumgarten (2015) ausführlich beschrieben wurden. Zum einen ist eine Abhängigkeit der Items zu erwarten, die zu einer Situation gehören. Darüber hinaus ist auch von einer Abhängigkeit der Items zwischen den Situationen auszugehen, da sich die Items auf eine Zielperson beziehen, über die im Laufe der Aufgabe immer mehr Wissen angeeignet wird (Baumgarten, 2015). Aus diesem Grund wurden in der vorliegenden Studie wie bei Conzelmann et al. (2013) sowie Baumgarten (2015) in den Messmodellen von SU ebenfalls Aufgaben-Parcels als Indikatoren verwendet. Zur Erstellung dieser wurden zunächst je SU-Aufgabe Itemscores entsprechend den Vorgaben von Süß et al. (2009) bestimmt, indem die (nicht standardisierten) quadrierten Distanzen zwischen den Antworten der Zielperson und den Antworten der Testpersonen ermittelt wurden. Diese wurden anschließend entsprechend der maximal möglichen Distanz gewichtet (Süß et al., 2009; vgl. auch Baumgarten, 2015). Danach wurde je Aufgabe und somit je Zielperson ein Gesamtscore gebildet, ebenfalls nach den Vorgaben von Süß et al. (2009). Hierfür wurden zunächst die zu einer Situation gehörigen Items zu Situationsscores zusammengefasst, indem die Wurzel aus der Summe der Itemscores gebildet wurde. Dieses Vorgehen wurde von den Autor:innen des MTSI gewählt, um die beschriebenen Abhängigkeiten zwischen den Items zu berücksichtigen (Baumgarten, 2015). Im Anschluss wurden die Situationsscores zu einem Gesamtscore je Zielperson aufsummiert, die schließlich als Indikatoren im Messmodell für SU verwendet wurden.

6.2.4.3 Zusammenhang PU und Selbst-Fremd Übereinstimmung. Die Bestimmung der Akkuratheit von Persönlichkeitsbeurteilungen mittels Selbst-Fremd Übereinstimmung wurde als klassische Operationalisierung einer Fähigkeit der Beurteiler:innen aufgefasst, sodass der Zusammenhang zu PU ebenfalls mit Hilfe eines zweifaktoriellen CFA-Modells

untersucht werden sollte. Als Indikatoren sollte ein Akkuratheits-Gesamtscore je Zielperson verwendet werden, da von einer Abhängigkeit der Beurteilungen, die zu einer Zielperson gehören, ausgegangen wurde.

Zur Bestimmung der Selbst-Fremd Übereinstimmung wurde ein anderes Vorgehen gewählt als in der bisherigen Forschung (vgl. Abschnitt 2.3). Dies hatte in erster Linie folgenden Grund: Da über die Zielpersonen der SU-Aufgaben nur sehr selektive Information zur Verfügung stand, wurde auf Basis des RAM (Funder, 1995) angenommen, dass sich nicht alle NEO-FFI Items für eine Fremdbeurteilung eignen. Es war zu erwarten, dass die Testpersonen bei einigen Items keine angemessene Fremdbeurteilung vornehmen konnten, da hierzu keine relevante Information über die Zielperson zur Verfügung stand. Diese Items wurden daher als ungeeignet angesehen, um interindividuelle Unterschiede in einer Fähigkeit der Beurteiler:innen zu erfassen (vgl. Rogers & Biesanz, 2019). Die ungeeigneten Items sollten über eine Itemselektion identifiziert und eliminiert werden. Da in den SU-Aufgaben je Zielperson unterschiedliche Information präsentiert wird (Süß et al., 2009), wurde zudem angenommen, dass je nach Zielperson unterschiedliche Items ungeeignet waren, sodass die Itemselektion für jede Zielperson separat erfolgte.

Die klassischen Ansätze zur Bestimmung der Akkuratheit unter Verwendung von Korrelations- und Regressionsanalysen resultieren in einem Gesamtscore je Testperson (z.B. Back & Nestler, 2016; vgl. Abschnitt 2.3), liefern aber keine Itemscores. Verwendet wurden in der vorliegenden Studie daher in Anlehnung an die Auswertung der AcquA-PU Aufgaben die standardisierten Distanzen zwischen NEO-FFI Fremdeinschätzung durch die Testperson und NEO-FFI Selbsteinschätzung der Zielperson. Die standardisierten Distanzwerte weisen große Ähnlichkeit zu den ansonsten oftmals verwendeten Profilkorrelationen auf, haben aber den Vorteil, dass sie die Bestimmung eines Akkuratheitswerts je Item erlauben. Wie von Legree et al. (2010) erläutert wurde, ist der Mittelwert aus den *quadrierten* standardisierten Distanzen, der über alle Items beispielsweise des NEO-FFI gebildet wird, äquivalent zur Korrelation zwischen den selbst- und fremdeingeschätzten NEO-FFI Items und beide Werte über eine lineare Transformation ineinander überführbar. Hinzu kommt, dass die standardisierten und die quadrierten standardisierten Distanzwerte eine fast perfekte Korrelation aufweisen (Legree et al., 2010), sodass die Verwendung der standardisierten Distanzen anstelle der Profilkorrelation angemessen erschien. Ein Nachteil der gewählten Methode bestand allerdings darin, dass eine Kontrolle der Stereotype Accuracy-Komponente, die in den standardisierten Distanzwerten ebenso enthalten sein müsste wie in den Profilkorrelationen (Funder, 1999; Furr, 2008; Kenny & Winquist, 2001), nicht möglich war,

unter anderem da nicht für alle Zielpersonen dieselben Items zur Bestimmung der Akkuratheit verwendet wurden.

Die Itemselektion der gebildeten Akkuratheitswerte erfolgte nicht über alle NEO-FFI Items einer Zielperson hinweg, sondern je Big Five-Faktor, um mögliche Abhängigkeiten der Akkuratheitswerte von Items eines Faktors zu berücksichtigen. Innerhalb einer Zielperson wurde die Selektion analog zu der Selektion bei den AcquA-PU Aufgaben vorgenommen. Im Anschluss wurden die selektieren Akkuratheitswerte eines Big Five-Faktors zusammengefasst, sodass je Zielperson fünf Big Five-Akkuratheitswerte resultierten. Vor der Zusammenfassung der Big Five-Akkuratheitswerte zu einen Akkuratheits-Gesamtscore je Zielperson wurden diese mit Hilfe eines einfaktoriellen CFA-Modells auf Eindimensionalität geprüft und gegebenenfalls weitere Selektionen vorgenommen. Die Akkuratheits-Gesamtscores sollten schließlich im oben beschriebenen zweifaktoriellen Modell zur Untersuchung des Zusammenhangs mit PU als Indikatoren verwendet werden.

6.2.4.4 Zusammenhang PU, schlussfolgerndes Denken und Merkfähigkeit. Um den Zusammenhang zwischen PU, schlussfolgerndem Denken und Merkfähigkeit zu untersuchen, wurde ein dreifaktorielles CFA-Modell spezifiziert. Als Indikatoren für das schlussfolgernde Denken wurden die Summenwerte der drei Untertests des Moduls 4 aus dem WIT-2 verwendet. Für Merkfähigkeit sollten ebenfalls drei Parcels als Indikatoren verwendet werden. In Anlehnung an die im Manual des WIT-2 geschätzten Modelle (vgl. Kersting et al., 2008) wurden drei Parcels gebildet, indem die ausgewerteten Items der verbalen, numerischen und figuralen Merkfähigkeit zu drei getrennten Summenscores zusammengefasst wurden.

6.2.4.5 Power. Um einschätzen zu können, ob für die geplanten Analysen genügend Testpersonen erhoben wurden, wurde auch in dieser Studie die Power im Zusammenhang mit dem Messmodell für PU bestimmt. Die Schätzung erfolgte unter Annahme eines eindimensionalen Messmodells mit fünf Indikatoren für den χ^2-Test des absoluten Fits mit H_0: RMSEA = 0, α = .05, angenommener Effekt: RMSEA = .10, df = 5 sowie einem N von 162 und wurde erneut mit Hilfe des von Preacher und Coffman (2006) publizierten R-Skript Generators vorbereitet. Die Poweranalyse in R ergab ein Ergebnis von 1-β = .57 und blieb damit deutlich unter der üblichen Mindestanforderung von .80, für die eine Stichprobengröße von mindestens 258 notwendig gewesen wäre.

Darüber hinaus erfolgte die Bestimmung der Power für die ebenfalls im Fokus der Studie stehenden Zusammenhänge von PU und SU, PU, schlussfolgerndem Denken und Merkfähigkeit sowie PU und der Selbst-Fremd Übereinstimmung. Durchgeführt wurde die Powerbestimmung mit Hilfe des Monte Carlo Ansatzes in Mplus (10000 Replikationen; Muthén & Muthén, 2002, 2017). Für die Schätzung wurden standardisierte Ladungen von .56 bei PU (durchschnittliche Ladung der AcquA-PU Aufgaben aus Studie 2 bei Verwendung des standardisierten Distanzscorings), .64 bei SU (durchschnittliche Ladung der drei Aufgaben bei Conzelmann et al., 2013, Studie 2), .74 bei schlussfolgerndem Denken und .70 bei Merkfähigkeit (durchschnittliche Ladung der Aufgaben im Manual des WIT-2; Kersting et al., 2008) sowie .40 bei der Selbst-Fremd Übereinstimmung (konservativere Annahme) angenommen. Auf Grund der in den Zielen formulieren Annahmen über die Zusammenhänge der Konstrukte wurden folgende latente Korrelationen in den Poweranalysen angenommen: 1.) PU und SU: .30, 2.) PU und schlussfolgerndes Denken: .50, 3.) PU und Merkfähigkeit: .30, 4.) PU und Selbst-Fremd Übereinstimmung: .30. Als *N* wurden die für die jeweilige Analyse zur Verfügung stehenden vollständigen Datensätze verwendet.

Für den Zusammenhang zwischen PU und SU ergab sich eine Power von .75 ($N = 157$), für den Zusammenhang zwischen PU und schlussfolgerndem Denken sowie PU und Merkfähigkeit ergab sich eine Power von .99 beziehungsweise .72 ($N = 129$; angenommener Zusammenhang zwischen logischem Schlussfolgern und Merkfähigkeit: .50) und für den Zusammenhang von PU und der Selbst-Fremd Übereinstimmung resultierte eine Power von .48 ($N = 159$; es traten zudem bei einigen Iterationen Schätzprobleme auf). Somit blieben auch hier einige Ergebnisse unter der üblichen Mindestanforderung. Eine Weiterführung der Erhebungen war aus Zeitgründen nicht möglich.

6.3 Ergebnisse

Die für die Analysen zur Verfügung stehende Stichprobengröße variierte in Abhängigkeit der betrachteten Instrumente zwischen $N = 128$ und 162. Dies hatte mehrere Gründe: Zum einen wurden nicht alle eingesetzten Instrumente von allen Testpersonen bearbeitet. Zum anderen traten bei drei Testpersonen technische Probleme bei der Bearbeitung der MTSI-3 Aufgaben zur Erfassung von SU auf, was zu einer größeren Anzahl an fehlenden Werten bei jeweils einer Aufgabe führte. Darüber hinaus konnte auf Grund von Problemen bei der Durchführung bei einer Testperson die NEO-FFI Fremdbeurteilung der Zielperson RF nicht erfolgen.

Die im Folgenden präsentierten Ergebnisse im Zusammenhang mit den AcquA-PU Aufgaben basieren auf der Anwendung des standardisierten Distanzscorings. Die Ergebnisse unter Verwendung des dichotomen Scorings befinden sich in Anhang E. Näher eingegangen wird auf diese nur vereinzelt und insbesondere bei relevanten Abweichungen zwischen beiden Scoring-Methoden.

6.3.1 AcquA-PU Aufgaben: Itemanalysen und Itemselektion

Die Itemselektion der AcquA-PU Aufgaben erfolgte wie in der vorherigen Studie. Schiefe, Kurtosis sowie Histogramme der Itemscores deuteten darauf hin, dass bei einer Vielzahl der Items keine univariate Normalverteilung vorlag (Range Schiefe: -2.05 bis 5.19; 21 der 70 Items zeigten eine Schiefe > |2|; Range Kurtosis: -1.41 bis 38.20), sodass bei keiner Aufgabe von einer multivariaten Normalverteilung ausgegangen werden konnte. Für die CFA-Modelle der Itemselektion wurde daher der MLM-Schätzer verwendet (Muthén & Muthén, 2017).

Bei Aufgabe PU1p wurden im Rahmen der Itemselektion alle Items außer jenen mit der korrekten Antwort $6.\overline{6}$ eliminiert. Diese vier Items zeigten in den im Anschluss an die Itemselektion durchgeführten Itemanalysen allerdings als einzige negative Trennschärfen zwischen -.48 und -.19. Auch das im Anschluss an die vollständige Itemselektion gebildete Aufgaben-Parcel zeigte im eindimensionalen Messmodell für PU eine negative standardisierte Ladung in Höhe von -.55. Da diese Ergebnisse klar darauf hindeuteten, dass ungeeignete Items selektiert und ein für die Operationalisierung von PU ungeeignetes Parcel gebildet wurde, wurde die Itemselektion bei Aufgabe PU1p wiederholt. Hierbei wurden zunächst die vier zuvor genannten Items eliminiert und im Anschluss die Itemselektion nach dem üblichen Vorgehen fortgesetzt. Bei dieser zweiten Itemselektion wurden für Aufgabe PU1p fünf Items ausgewählt, für die sich positive Trennschärfen ergaben und auch das gebildete Aufgaben-Parcel zeigte im späteren Messmodell eine positive standardisierte Ladung (s. unten), sodass die folgenden Ergebnisse auf der zweiten Itemselektion basieren. Anzumerken ist hierbei, dass alle fünf selektierten Items der Aufgabe PU1p die korrekte Antwort 0 aufweisen. Es wurden folglich alle Items, die den zwei modellierten Reaktionen der probabilistischen Kontingenzen entsprechen (Reaktion A und B; d.h. alle Items mit den korrekten Antworten $6.\overline{6}$ und $3.\overline{3}$), eliminiert. Dieses Ergebnis zeigte sich auch bei Verwendung des dichotomen Scorings. Hier fiel im Rahmen der Itemselektion zudem auf, dass alle Items mit der korrekten Antwort $6.\overline{6}$ auf Grund eines Mittelwerts > .95 eliminiert wurden.

Insgesamt wurde beim standardisierten Distanzscoring die Itemanzahl von ursprünglich 70 auf 35 reduziert. Deskriptive Statistiken der selektierten Items sind Tabelle F1 in Anhang F

zu entnehmen. Die part-whole korrigierten Trennschärfen dieser Items, berechnet über alle selektierten Items hinweg, liegen bei .19 bis .63 (*M* = .41, *SD* = .13) und die Itemmittelwerte bei 0.33 bis 1.27 (*M* = 0.50, *SD* = .19). Zu beachten ist hierbei, dass auf Grund der standardisierten Distanzen höhere Itemmittelwerte eine höhere Schwierigkeit anzeigen. Die deskriptiven Statistiken der im Anschluss an die Itemselektion gebildeten aufgabenweisen Parcels und eines PU-Gesamtscores finden sich in Tabelle 6.5. Die Erstellung der Parcels und des Gesamtscores erfolgte durch Mittelwertbildung.

Tabelle 6.5

Deskriptive Statistiken der Personality Understanding (PU)-Parcels sowie des gesamten AcquA-PU Tests (PU_{ges}) nach der Itemselektion (standardisiertes Distanzscoring)

Parcel	Itemanzahl[a]	*M*	*SD*	Schiefe[b]	Kurtosis[c]
PU1p	5 (12)	0.40	0.23	1.27	3.51
PU2	6 (12)	0.49	0.29	1.20	1.68
PU3	12 (24)	0.44	0.20	0.15	-0.03
PU5k	7 (10)	0.70	0.27	-0.21	-0.35
PU6	5 (12)	0.47	0.26	0.79	1.19
PU_{ges}	35 (70)	0.50	0.18	-0.02	0.32

Anmerkungen. *N* = 162. Die verwendete Statistik zur Schätzung der Kurtosis nimmt beim Vorliegen einer Normalverteilung den Wert 0 an.

[a] In Klammern ist die Anzahl vor der Itemselektion angegeben. [b] *SE* = 0.19. [c] *SE* = 0.38.

Ein Vergleich des Mittelwerts von Aufgabe PU1p (*M* = 0.40) mit dem Mittelwert von Aufgabe PU1 aus Studie 2 (*M* = 0.44; vgl. Tabelle 5.3) zeigt für PU1p keinen höheren Mittelwert und somit keine höhere Schwierigkeit. Zudem ist Aufgabe PU1p deskriptiv betrachtet die einfachste der vorliegenden Studie. Aufgabe PU5k lässt sich nicht sinnvoll mit Aufgabe PU5 aus Studie 2 vergleichen. Allerdings resultierte für diese Aufgabe deskriptiv der größte Mittelwert aller Aufgaben der vorliegenden Studie und somit die höchste Schwierigkeit. Auch über alle drei Studien hinweg betrachtet, weist PU5k die deskriptiv höchste Schwierigkeit auf (PU5 aus Studie 2 ausgenommen). Beim dichotomen Scoring resultierte für Aufgabe PU1p ein ähnliches Bild, während Aufgabe PU5k bei dieser Scoring-Methode nicht mehr die deskriptiv höchste Schwierigkeit zeigte (vgl. Abschnitt E.1 in Anhang E).

6.3.2 AcquA-PU Aufgaben: Messmodell und Reliabilität

Schiefe, Kurtosis sowie Histogramme deuteten bei einigen PU-Parcels auf Abweichungen von einer univariaten Normalverteilung hin. Die Ergebnisse des Shapiro-Wilk-Tests auf Normalverteilung (vgl. Tabelle F2 in Anhang F) bestätigten dies für drei Parcels, sodass für das einfaktorielle Messmodell für PU der robuste MLM-Schätzer verwendet wurde. Das angenommene Modell zeigte mit $\chi^2 = 4.76$, $df = 5$, $p = .45$, CFI = 1, RMSEA = 0, 90% CI = [.00, .11] einen perfekten Fit ($N = 162$; standardisierte Faktorladungen siehe Tabelle 6.6). Die Schätzung der Reliabilität erfolgte auf Grund der bestätigten Eindimensionalität unter Verwendung von Cronbachs α sowie McDonalds ω auf Basis der fünf PU-Parcels: α = .79, 95% CI = [.72, .84]; ω = .79, 95% CI = [.73, .84].

Tabelle 6.6

Ergebnis der konfirmatorischen Faktorenanalyse der Personality Understanding (PU)-Parcels (standardisiertes Distanzscoring)

Parcel	standardisierte Faktorladung
PU1p	.62***
PU2	.65***
PU3	.82***
PU5k	.59***
PU6	.65***

Anmerkungen. $N = 162$.

***$p < .001$.

Unter Verwendung des dichotomen Scorings ergaben sich sowohl hinsichtlich der standardisierten Ladungen im Messmodell als auch bei den Reliabilitätsschätzungen nennenswerte Unterschiede (vgl. Abschnitt E.2 in Anhang E). Die standardisierten Ladungen fielen mit .26 (PU2) bis .61 (PU3) für alle Aufgaben geringer aus als beim standardisierten Distanzscoring und für die Reliabilitätsschätzungen zeigten sich mit α = .54 und ω = .55 noch deutlichere Unterschiede.

6.3.3 MTSI-3 SU

6.3.3.1 Umgang mit fehlenden Werten. Wie bereits beschrieben, traten bei den SU-Aufgaben des MTSI-3 bei drei Testpersonen technische Probleme auf, was zu einer größeren

Anzahl fehlender Werte führte (33 bis 39 % fehlende Werte bei jeweils einer Zielperson). Zudem konnte bei zwei Testpersonen der zweite Erhebungstermin nicht stattfinden, sodass nur Daten für jeweils eine Zielperson vorlagen. Diese fünf Testpersonen wurden auf Grund der hohen Anteile fehlender Werte von den weiteren Analysen ausgeschlossen.

Da die Testpersonen bei den SU-Aufgaben einzelne Items überspringen konnten, wurde in einem nächsten Schritt der Anteil fehlender Werte der übrigen 157 Testpersonen je Zielperson sowie die fehlenden Antworten je Item betrachtet. Für die Zielperson RF lag die Anzahl fehlender Werte bei 0 bis 3 (7 %) nicht beantworteten Items pro Testperson und bei 0 bis 2 (1 %) fehlenden Antworten pro Item. Bei FB lag die Anzahl fehlender Werte bei 0 bis 4 (11 %) nicht beantworteten Items pro Testperson und bei 0 bis 4 (3 %) fehlenden Antworten pro Item. Für HR lag die Anzahl fehlender Werte bei Betrachtung der Testpersonen bei 0 bis 6 (10 %) nicht beantworteten Items und bei Betrachtung der Items bei 0 bis 9 (6 %) fehlenden Antworten. Über alle Testpersonen und Items hinweg lag der Anteil fehlender Werte bei 0.3 % (RF), 0.3 % (FB) und 0.6 % (HR). Eine Inspektion der Muster der fehlenden Werte ergab, dass diese in der Regel bei Items auftraten, die zu derselben Situation gehören und zusammen auf einer Monitorseite präsentiert wurden, was darauf hindeutet, dass diese Items versehentlich übersprungen wurden.

Die fehlenden Werte wurden je Testperson mittels Expectation-Maximization Algorithmus auf Basis aller anderen gescorten und bereits gewichteten Items der entsprechenden Zielperson ersetzt. Zudem wurden imputierte Werte, die außerhalb des theoretisch möglichen Wertebereichs lagen, durch den kleinsten beziehungsweise größten möglichen Wert ausgetauscht.

6.3.3.2 Itemanalysen und Reliabilität. Die Reliabilität der SU-Aufgaben wurde auf Basis der Gesamtscores der drei Zielpersonen bestimmt und fiel mit Cronbachs $\alpha = .56$, 95% CI = [.44, .66] und McDonalds $\omega = .57$ [.46, .69] eher gering aus. Bei der Schätzung des Konfidenzintervalls für ω traten zudem wiederholt Schätzprobleme in Form negativer Varianzen auf. Die deskriptiven Statistiken der gebildeten Gesamtscores sind in Tabelle 6.7 zu finden. Zusätzlich werden hier auch Schätzungen von Cronbachs α der einzelnen Zielpersonen auf Basis der Situationsscores (Bestimmung in Anlehnung an Baumgarten, 2015) angegeben, um einen Vergleich mit Ergebnissen bisheriger Studien zu ermöglichen.

Tabelle 6.7

Deskriptive Statistiken der Gesamtscores der Zielpersonen RF, FB und HR des MTSI-3, Anzahl der Situationsscores, die in die Berechnung des Gesamtscores eingingen, sowie Reliabilitätsschätzungen (Cronbachs α) auf Basis der Situationsscores

Zielperson	Anzahl Situationen	*M*	*SD*	Schiefe[a]	Kurtosis[b]	α [95% CI]
RF	11	46.98	6.73	0.29	-0.54	.37 [.23, .48]
FB	8	35.05	5.77	0.60	0.23	.47 [.32, .57]
HR	9	54.89	8.46	0.42	0.29	.70 [.62, .77]

Anmerkungen. N = 157. CI = Konfidenzintervall. Die verwendete Statistik zur Schätzung der Kurtosis nimmt beim Vorliegen einer Normalverteilung den Wert 0 an.

[a] SE = 0.19. [b] SE = 0.38.

6.3.4 Zusammenhang PU und SU

Zur Schätzung des Zusammenhangs zwischen PU und SU auf Basis des zweifaktoriellen CFA-Modells wurde auf Grund der Verteilung der PU-Parcels erneut der MLM-Schätzer verwendet. Das Modell zeigte mit χ^2 = 18.34, df = 19, p = .50, CFI = 1, RMSEA = 0, 90% CI = [.00, .07] eine perfekte Passung auf die Daten (N = 157). Die standardisierten Faktorladungen sowie die Faktorkorrelation sind in Tabelle 6.8 zu finden.

Für PU und SU resultierte mit -.22 ein nicht signifikanter latenter Zusammenhang in unerwarteter Richtung. Auf Grund der bei beiden Konstrukten verwendeten Distanzwerte, bei denen ein niedrigerer Wert eine bessere Leistung anzeigt, deutet diese latente Korrelation darauf hin, dass eine bessere Leistung bei PU tendenziell mit einer schlechteren Leistung bei SU einhergeht. Auch unter Verwendung des dichotomen Scorings bei den AcquA-PU Aufgaben zeigte sich ein Zusammenhang zwischen PU und SU in unerwarteter Richtung, wobei dieser mit .12 (n.s.) etwas schwächer ausfiel (vgl. Anhang E Tabelle E3).

Tabelle 6.8

Ergebnis der konfirmatorischen Faktorenanalyse der Personality Understanding (PU)-Parcels (standardisierte Distanzen) und Social Understanding (SU)-Parcels

	standardisierte Faktorladung	
Parcel	PU	SU
PU1p	.62***	
PU2	.66***	
PU3	.83***	
PU5k	.60***	
PU6	.65***	
RF		.46***
FB		.62***
HR		.60***
Faktorkorrelation [95% CI]		
PU - SU	-.22 [-.46, .03]	

Anmerkungen. $N = 157$. CI = Konfidenzintervall.

***$p < .001$.

Auf Grund dieses unerwarteten Ergebnisses wurden in einem nächsten Schritt die SU-Aufgaben näher betrachtet. Hierfür wurde der latente Zusammenhang zwischen SU und schlussfolgerndem Denken sowie Merkfähigkeit mit Hilfe eines dreifaktoriellen CFA-Modells exploriert ($N = 128$; für deskriptive Statistiken und Reliabilität des schlussfolgernden Denkens und der Merkfähigkeit siehe Abschnitt 6.3.7). Das Ziel bestand darin, die geschätzten Zusammenhänge im Rahmen der Interpretation mit den Ergebnissen bisheriger Studien zum MTSI zu vergleichen, um die Validität der verwendeten SU-Aufgaben besser einschätzen zu können. Es zeigte sich eine latente Korrelation zwischen SU und schlussfolgerndem Denken von -.29 und eine latente Korrelation zwischen SU und Merkfähigkeit von -.24 bis -.22 (vgl. ausführliche Ergebnisse in Tabelle F3 in Anhang F). Unter Berücksichtigung der verwendeten Scoring-Methoden bei den SU-Aufgaben sowie Aufgaben des WIT-2 deuten die geschätzten Zusammenhänge darauf hin, dass eine bessere Leistung bei SU tendenziell mit einer besseren Fähigkeit zum schlussfolgernden Denken sowie einer besseren Merkfähigkeit einhergeht.

Darüber hinaus wurde exploriert, ob sich der Zusammenhang zwischen PU und SU ändert, wenn dieser getrennt für die drei Zielpersonen bestimmt wird. Hierfür wurde je

Zielperson ein zweifaktorielles CFA-Modell geschätzt (N = 157; MLM-Schätzer). Als Indikatoren für PU wurden die fünf Aufgaben-Parcels verwendet und als Indikatoren für SU, erfasst auf Basis einer der drei Zielpersonen, die Situationsscores der jeweiligen Zielperson. Beim standardisierten Distanzscoring resultierten nicht signifikante latente Korrelationen zwischen PU und SU von -.02 (RF), -.19 (FB) und -.23 (HR). Für das dichotome Scoring zeigten sich latente Korrelationen von .06 (RF), .13 (FB) und .10 (HR). Die Fits der Modelle waren, mit einer Ausnahme, in der der CFI mit .94 knapp unter der Grenze für einen guten Fit von .95 blieb (vgl. Hu & Bentler, 1999), gut bis perfekt. Mit Ausnahme von Zielperson RF beim standardisierten Distanzscoring zeigten sich hier also kaum Unterschiede zu den Ergebnissen in Tabelle 6.8 sowie Tabelle E3 in Anhang E .

6.3.5 Selbst-Fremd Übereinstimmung: Itemselektion und Messmodell

Als nächstes folgten die Vorbereitungen für die Untersuchung des Zusammenhangs zwischen den AcquA-PU Aufgaben und der Übereinstimmung zwischen NEO-FFI Selbstbericht der Zielpersonen RF, FB und HR und NEO-FFI Fremdbericht der Testpersonen. Die in Abschnitt 6.2.4.3 beschriebene Itemselektion der Akkuratheitswerte je NEO-FFI Item führte zur Elimination von 21 (RF), 19 (FB) und 17 (HR) der ursprünglich 60 Akkuratheitswerte. Die im Anschluss gebildeten Akkuratheitswerte je Big Five-Faktor wurden in einem nächsten Schritt wiederum auf Eindimensionalität geprüft, indem je Zielperson ein einfaktoriellen CFA-Modell geschätzt wurde. Hierbei zeigte sich bei der Zielperson RF, dass der Wert für den Faktor Neurotizismus eine negative standardisierte Ladung von -.40 aufwies. Da die Big Five-Akkuratheitswerte eine Leistung erfassen sollten, wurde der Wert für Neurotizismus auf Grund des negativen Vorzeichens eliminiert. Das resultierende Messmodell auf Basis der vier übrigen Big Five-Akkuratheitswerte wies mit $\chi^2 = 4.21$, $df = 2$, $p = .12$, CFI = .93, RMSEA = .08, 90% CI = [.00, .20] (N = 159, MLM-Schätzer, standardisierte Faktorladungen: .21 bis .76) entsprechend den Kriterien für CFI und RMSEA von Hu und Bentler (1999) beziehungsweise Browne und Cudeck (1992) keinen guten Fit auf. Eine weitere Selektion wurde auf Grund der wenigen Indikatoren nicht vorgenommen. Cronbachs α auf Basis der vier Big Five-Akkuratheitswerte ergab einen Wert von .54, 95% CI = [.40; .64] und ω_H lag bei .57, 95% CI = [.46, .76][17]. Bei Zielperson FB zeigten alle fünf Big Five-Akkuratheitswerte positive Faktorladungen (standardisierte Ladungen: .25 bis .79), ein guter Fit entsprechend der zuvor zitierten Kriterien konnte allerdings erst nach Elimination des Wertes für Neurotizismus

[17] Auf Grund der nicht vorliegenden klaren Eindimensionalität wurde ω_H anstelle von ω bestimmt (siehe Kelley & Pornprasertmanit, 2016).

erreicht werden (χ^2 = 1.71, *df* = 2, *p* = .42, CFI = 1, RMSEA = 0, 90% CI = [.00, .15], *N* = 160, MLM-Schätzer, standardisierte Ladungen: .25 bis .85; α = .62, 95% CI = [.51; .70], ω = .66, 95% CI = [.59, .74]). Im Modell der Zielperson HR ergab sich für den Akkuratheitswert der Gewissenhaftigkeit eine hohe negative standardisierte Ladung von -.61. Der Modellfit nach Elimination dieses Indikators ergab, je nachdem welcher Fit-Index betrachtet wurde, einen akzeptablen bis guten Fit (χ^2= 3.64, *df* = 2, *p* = .16, CFI = .98, RMSEA = .07, 90% CI = [.00, .19], *N* = 162, MLM-Schätzer, standardisierte Ladungen: .25 bis .74, α = .64, 95% CI = [.56, .70], ω_H = .71, 95% CI = [.63, .78]). Eine weitere Selektion wurde auf Grund der Indikatoranzahl nicht vorgenommen. Im Zusammenhang mit allen Analysen muss zudem darauf hingewiesen werden, dass bei der Schätzung der Konfidenzintervalle Probleme auftraten (insb. negative Varianzen).

Im nächsten Schritt wurde die Eindimensionalität der Akkuratheits-Gesamtscores der Zielpersonen überprüft. Die Verwendung eines einfaktoriellen Modells mit den drei Gesamtscores als Indikatoren war auf Grund der Indikatoranzahl nicht sinnvoll. Daher wurde ein Higher-Order Modell mit drei Faktoren erster Ordnung für die drei Zielpersonen sowie einem Faktor höherer Ordnung geschätzt. Die zuvor gebildeten und selektierten Big Five-Akkuratheitswerte je Zielperson dienten als Indikatoren der Faktoren erster Ordnung. Dieses Modell zeigte mit χ^2 = 61.48, *df* = 51, *p* = .15, CFI = .96, RMSEA = .04, 90% CI = [.00, .07] einen guten Fit (*N* = 159, MLM-Schätzer). Wie Tabelle 6.9 zu entnehmen ist, fielen die standardisierten Ladungen auf dem Faktor höherer Ordnung mit .19, .36 und .96 sehr variabel und allesamt nicht signifikant aus. Dies deutete darauf hin, dass die Faktoren erster Ordnung keine substantiellen Zusammenhänge aufweisen, die durch den Faktor höherer Ordnung erklärt werden können. Zusätzlich bestätigt wurde dies durch die Schätzung eines Modells, in dem anstelle des Faktors höherer Ordnung die latenten Korrelationen zwischen den Faktoren erster Ordnung spezifiziert wurden (Fit siehe oben). Die latenten Korrelationen lagen bei .07 (RF und FB; 95% CI = [-.13, .27]), .18 (RF und HR; 95% CI = [-.02, .38]) sowie .35 (FB und HR; 95% CI = [.15, .54]). Es musste für die folgenden Analysen somit angenommen werden, dass die Akkuratheits-Gesamtscores der Zielpersonen keine Eindimensionalität aufweisen und somit der Faktor höherer Ordnung keine einheitliche Fähigkeit der Beurteiler:innen abbildet.

Tabelle 6.9

Standardisierte Faktorladungen der Big Five-Akkuratheitswerte der Zielpersonen RF, FB und HR sowie der Faktoren erster Ordnung im Higher-Order Modell

	Faktoren erster Ordnung		
Parcel	RF	FB	HR
RF_e	.86***		
RF_o	.18*		
RF_a	.58***		
RF_c	.27*		
FB_e		.70***	
FB_o		.39***	
FB_a		.83***	
FB_c		.26***	
HR_n			.25**
HR_e			.75***
HR_o			.62***
HR_a			.63***
Faktor erster Ordnung	Faktor höherer Ordnung		
RF	.19		
FB	.36		
HR	.96		

Anmerkungen. N = 159. n = Neurotizismus; e = Extraversion; o = Offenheit für Erfahrungen; a = Verträglichkeit; c = Gewissenhaftigkeit.

*p < .05. **p < .01. ***p < .001.

6.3.6 Zusammenhang PU und Selbst-Fremd Übereinstimmung

Auf Grund der fehlenden Zusammenhänge zwischen den Akkuratheitswerten der Zielpersonen wurde der Zusammenhang mit PU, anders als ursprünglich geplant, getrennt für die Zielpersonen geschätzt. Hierfür wurde ein vierfaktorielles Modell mit einem Faktor für PU und drei Faktoren für die drei Zielpersonen geschätzt (vgl. Tabelle 6.10). Bei der Interpretation der latenten Korrelationen ist zu beachten, dass sowohl bei der Auswertung der AcquA-PU Aufgaben, als auch bei der Bildung der Akkuratheitswerte standardisierte Distanzwerte verwendet wurden, also in beiden Fällen niedrigere Werte eine bessere Leistung anzeigen.

Tabelle 6.10

Ergebnis der konfirmatorischen Faktorenanalyse der Personality Understanding (PU)-Parcels (standardisierte Distanz) und Big Five-Akkuratheitswerte der Zielpersonen RF, FB und HR

	standardisierte Faktorladung			
Parcel/Faktor	PU	RF	FB	HR
PU1p	.63***			
PU2	.66***			
PU3	.83***			
PU5k	.59***			
PU6	.64***			
RF_e		.84***		
RF_o		.18*		
RF_a		.59***		
RF_c		.27*		
FB_e			.74***	
FB_o			.39***	
FB_a			.79***	
FB_c			.26**	
HR_n				.24**
HR_e				.75***
HR_o				.63***
HR_a				.63***
Faktorkorrelationen [95% CI]				
RF	.16 [-.01, .32]			
FB	.11 [-.10, .31]	.07 [-.14, .27]		
HR	-.15 [-.34, .04]	.18 [-.01, .37]	.36 [.17, .54]	

Anmerkungen. N = 159. n = Neurotizismus; e = Extraversion; o = Offenheit für Erfahrungen; a = Verträglichkeit; c = Gewissenhaftigkeit; CI = Konfidenzintervall. Modell-Fit (MLM-Schätzer): χ^2 = 126.85, df = 113, p = .18, CFI = .97, RMSEA = .03, 90% CI = [.00, .05].
*p < .05. **p < .01. ***p < .001.

Zwischen PU und der akkuraten Beurteilung von RF sowie FB zeigten sich kleine, aber nicht signifikante Zusammenhänge in erwarteter Richtung. Der Zusammenhang zwischen PU und der Akkuratheit bei HR fiel ebenfalls nicht signifikant, aber negativ aus, was als nicht erwartungskonform angesehen werden muss, wenn man davon ausgeht, dass in beiden Fällen Fähigkeiten erfasst werden sollten. Die Ergebnisse unter Verwendung des dichotomen Scorings bei PU, die vergleichbar ausfielen, sind in Tabelle E4 in Anhang E dargestellt.

6.3.7 WIT-2: Deskriptive Statistiken und Reliabilität

Deskriptive Statistiken und Schätzungen der Reliabilität der WIT-2 Skalen schlussfolgerndes Denken und Merkfähigkeit sowie deskriptive Statistiken der Untertests des schlussfolgernden Denkens finden sich in Tabelle 6.11. Die Auswertung erfolgte nach den Angaben des Manuals, indem Summenwerte der richtig gelösten Items gebildet wurden (Kersting et al., 2008). Die Reliabilitätsschätzungen erfolgten einerseits zum Vergleich mit den Ergebnissen des Manuals auf Itemebene durch Cronbachs α und andererseits auf Parcelebene durch McDonalds ω auf Basis der Parcels, die auch in den anschließenden Strukturanalysen verwendet wurden.

Tabelle 6.11

*Itemanzahl (*k*), deskriptive Statistiken und Reliabilität (Cronbachs α, McDonalds ω) der Skalen schlussfolgerndes Denken (R) und Merkfähigkeit (M) des WIT-2 sowie deskriptive Statistiken der Untertests Abwicklungen (Aw), Analogien (Al) und Zahlenreihen (Zn)*

Skala	*k*	*M*	*SD*	Schiefe[a]	Kurtosis[b]	α [95% CI][c]	ω [95% CI][d]
R	60	30.60	10.12	0.30	-0.54	.90 [.87, .92]	.72 [.63, .79]
Aw	20	10.61	4.18	0.10	-0.86		
Al	20	10.36	4.00	0.16	-0.61		
Zn	20	9.63	4.46	0.18	-0.22		
M	21	11.91	3.35	0.01	-0.48	.68 [.60, .74]	.73 [.66, .79][e]

Anmerkungen. N = 129. CI = Konfidenzintervall. Die verwendete Statistik zur Schätzung der Kurtosis nimmt beim Vorliegen einer Normalverteilung den Wert 0 an.

[a] *SE* = 0.21. [b] *SE* = 0.42. [c] Schätzung auf Itemebene. [d] Schätzung auf Parcelebene (drei Parcels). [e] Ergebnis auf Basis von Parcel-Set 1. Parcel-Set 2: ω = .72, 95% CI = [.65, .79]. Parcel-Set 3: ω = .75, 95% CI = [.68; .81]. Bei allen drei Parcel-Sets kam es bei Schätzung der Konfidenzintervalle wiederholt zu Schätzproblemen (negative Varianzen).

Cronbachs α fiel für beide Skalen geringer aus als im WIT-2 Manual, wobei die Unterschiede bei der Merkfähigkeit (α = .68 vs. .78 im Manual) größer ausfielen als beim schlussfolgernden Denken (α = .90 vs. .94 [stratifiziertes α] im Manual; Kersting et al., 2008). Beim schlussfolgernden Denken fiel die Schätzung der Reliabilität auf Parcelebene zudem geringer aus als auf Itemebene (ω = .72 vs. α = .90). Eine Überprüfung des Messmodells für Merkfähigkeit auf Basis der drei Merkfähigkeits-Parcels ergab eine standardisierte Ladung der verbalen Merkfähigkeit > 1. Aus diesem Grund wurden die ursprüngliche Parcelbildung verworfen und neue Parcels gebildet, indem die 21 Items zufällig und gleichmäßig auf drei Parcels aufgeteilt wurden. Zur Überprüfung der Stabilität der Ergebnisse wurden auf diese Weise insgesamt drei Parcel-Sets gebildet und jeweils die Reliabilität mittels McDonalds ω geschätzt. Die Ergebnisse aller drei Parcel-Sets sind in Tabelle 6.11 zu finden und zeigen konsistente Ergebnisse, die etwas höher ausfielen als die Ergebnisse auf Itemebene (ω = .72 bis .75 vs. α = .68). Die drei zufällig gebildeten Parcel-Sets wurden auch in den anschließenden Strukturanalysen verwendet.

Im Hinblick auf einen späteren Vergleich der Ergebnisse aus Studie 2 und Studie 3 wurde zudem exploriert, ob sich die Mittelwerte für das schlussfolgernde Denken der beiden Stichproben signifikant voneinander unterscheiden (Studie 2: M = 32.02, SD = 10.23; Studie 3: M = 30.60, SD = 10.12; vgl. Tabelle 5.6 und Tabelle 6.11). Ein Zweistichproben-t-Test für unabhängige Stichproben und homogene Varianzen ergab keinen signifikanten Unterschied: $t(331) = 1.25$, $p = .21$, $d = .14$.

6.3.8 Zusammenhang PU, schussfolgerndes Denken und Merkfähigkeit

Entsprechend des zweiten Studienziels wurde schließlich der Zusammenhang zwischen PU und schlussfolgerndem Denken sowie Merkfähigkeit bestimmt. Für das gemeinsame CFA-Modell wurde auf Grund der Verteilungen der PU-Parcels der MLM-Schätzer verwendet.

Unter Verwendung von Parcel-Set 1 der Merkfähigkeit zeigte das Modell eine gute Passung auf die Daten (N = 129): $\chi^2 = 47.31$, $df = 41$, $p = .23$, CFI = .98, RMSEA = .04, 90% CI = [.00, .07]. Die standardisierten Faktorladungen sowie die Faktorkorrelationen sind in Tabelle 6.12 zu finden. Bei der Interpretation der Faktorkorrelationen ist zu beachten, dass bei PU auf Grund des standardisierten Distanzscorings ein niedrigerer Wert eine bessere Leistung anzeigt. Die Ergebnisse deuten somit darauf hin, dass eine höhere Ausprägung auf dem Faktor PU tendenziell mit einer höheren Fähigkeit zum schlussfolgernden Denken und einer höheren Merkfähigkeit einhergeht. Der latente Zusammenhang zwischen PU und schlussfolgerndem Denken fiel mit -.40 zudem größer aus als der zwischen PU und Merkfähigkeit (-.21). Unter

Verwendung der Parcel-Sets 2 und 3 der Merkfähigkeit ergaben sich sehr ähnliche Ergebnisse, die in Tabellen F4 und F5 in Anhang F zu finden sind.

Tabelle 6.12

Ergebnis der konfirmatorischen Faktorenanalyse der Personality Understanding (PU)-Parcels (standardisierte Distanzen), Parcels des schlussfolgernden Denkens (R) und der Merkfähigkeit (M; Parcel-Set 1)

	standardisierte Faktorladung		
Parcel/Faktor	PU	R	M
PU1p	.67***		
PU2	.62***		
PU3	.77***		
PU5k	.56***		
PU6	.65***		
Al		.74***	
Zn		.62***	
Aw		.68***	
M1a			.52***
M1b			.79***
M1c			.74***
Faktorkorrelationen [95% CI]			
R	-.40 [-.58, -.23]		
M	-.21 [-.42, -.00]	.48 [.27, .69]	

Anmerkungen. N = 129. Al = Analogien; Zn = Zahlenreihen; Aw = Abwicklungen; M1a = Merkfähigkeit Parcel-Set 1, Parcel 1; M1b = Merkfähigkeit Parcel-Set 1, Parcel 2; M1c = Merkfähigkeit Parcel-Set 1, Parcel 3; CI = Konfidenzintervall.

***$p < .001$.

Zur Ergänzung der Ergebnisse wurde zudem eine latente Regression von PU auf schlussfolgerndes Denken und Merkfähigkeit durchgeführt. Der latente Regressionskoeffizient des schlussfolgernden Denkens lag bei -.43, 95% CI [-.69, -.16] und der der Merkfähigkeit bei -.03, 95% CI [-.28, .23] (Parcel-Set 1). Die latente Korrelation zwischen schlussfolgerndem

Denken und Merkfähigkeit sowie Modellfit und standardisierte Ladungen blieben im Vergleich zu den Ergebnissen des dreifaktoriellen Modells mit korrelierten Faktoren erwartungsgemäß unverändert. Unter Verwendung der Parcel-Sets 2 und 3 der Merkfähigkeit resultierten fast identische Ergebnisse: Bei Parcel-Set 2 ergaben sich latente Regressionskoeffizienten von -.43, 95% CI [-.69, -.18] für schlussfolgerndes Denken und -.02, 95% CI [-.27, .24] für Merkfähigkeit; bei Parcel-Set 3 ergaben sich Werte von -.43, 95% CI [-.69, -.17] für schlussfolgerndes Denken und -.02, 95% CI [-.26, .23] für Merkfähigkeit.

Die Ergebnisse unter Verwendung des dichotomen Scorings finden sich in Anhang E (Tabellen E5 bis E7). Hier resultierten für den latenten Zusammenhang zwischen PU und schlussfolgerndem Denken ähnlich hohe Schätzungen (r = .37 beim dichotomen Scoring vs. r = -.40 beim standardisierten Distanzscoring; Vorzeichen auf Grund der unterschiedlichen Scoring-Methoden erwartungskonform). Lediglich der Zusammenhang zwischen PU und Merkfähigkeit fiel bei Verwendung des dichotomen Scorings etwas höher aus (r = .30 bis .32 beim dichotomen Scoring vs. r = -.21 beim standardisierten Distanzscoring). Auch beim dichotomen Scoring wurde zusätzlich eine latente Regression durchgeführt. Der latente Regressionskoeffizient von schlussfolgerndem Denken fiel mit .30, 95% CI [-.02, .63] (Parcel-Set 1) geringer aus als beim standardisierten Distanzscoring. Der Koeffizient der Merkfähigkeit lag bei .21, 95% CI [-.10, .51] und fiel damit deutlich größer aus als beim standardisierten Distanzscoring (Parcel-Set 1[18]).

Der latente Zusammenhang zwischen PU und schlussfolgerndem Denken fiel in der vorliegenden Studie 3 größer aus als in Studie 2, insbesondere unter Verwendung des dichotomen Scorings (r = -.40 in Studie 3 vs. r = -.22 in Studie 2 [standardisierte Distanzen]; r = .37 in Studie 3 vs. r = .09 in Studie 2 [dichotomes Scoring]). Die Vergleichbarkeit der Ergebnisse ist allerdings eingeschränkt, da in Studie 3 die Aufgaben PU1 und PU5 in einer etwas anderen Form und Aufgabe PU4 gar nicht verwendet wurden. Um die Vergleichbarkeit zwischen den beiden Studien zu erhöhen, wurde zusätzlich ein zweifaktorielles Modell für den Zusammenhang zwischen PU und schlussfolgerndem Denken geschätzt, in dem als Indikatoren für PU nur diejenigen Aufgaben verwendet wurden, die in beiden Studien in identischer Form eingesetzt wurden (PU2, PU3 sowie PU6). Dieses Modell wurde auf Basis der Daten aus beiden Studien geschätzt und die Ergebnisse anschließend miteinander verglichen. Verwendet

[18] Für Parcel-Set 2 ergaben sich latente Regressionskoeffizienten von .32, 95% CI [-.00, .64] für schlussfolgerndes Denken und .16, 95% CI [-.16, .48] für Merkfähigkeit. Für Parcel-Set 3 ergaben sich Koeffizienten von .31, 95% CI [-.02, .63] für schlussfolgerndes Denken und .18, 95% CI [-.12, .49] für Merkfähigkeit.

wurde erneut der MLM-Schätzer. Die Ergebnisse unter Verwendung des standardisierten Distanzscorings finden sich in Tabelle 6.13, die unter Verwendung des dichotomen Scorings in Tabelle E8 in Anhang E.

Tabelle 6.13

Standardisierte Faktorladungen der Personality Understanding (PU)-Parcels 2, 3 und 6 (standardisierte Distanzen) und Parcels des schlussfolgernden Denkens (R) sowie Faktorkorrelation zwischen PU und R auf Basis der Daten aus Studie 2 und 3

	Studie 2		Studie 3	
Parcel/Faktor	PU	R	PU	R
PU2	.49***		.64***	
PU3	.71***		.79***	
PU6	.42***		.61***	
Al		.62***		.79***
Zn		.65***		.58***
Aw		.76***		.66***
Faktorkorrelation [95% CI]				
PU - R	-.25 [-.48, -.01]		-.39 [-.57, -.21]	
N	204		129	
χ^2	3.12		7.81	
df	8		8	
p	.93		.45	
CFI	1		1	
RMSEA [90% CI]	.00 [.00, .03]		.00 [.00, .10]	

Anmerkungen. Al = Analogien; Zn = Zahlenreihen; Aw = Abwicklungen. CI = Konfidenzintervall.

***$p < .001$.

Unter Verwendung beider Scoring-Methoden fiel der latente Zusammenhang zwischen PU und schlussfolgerndem Denken in Studie 3 weiterhin größer aus als in Studie 2. Verglichen mit den vorherigen Ergebnissen unter Verwendung von mehr Indikatoren für PU kam es bei Verwendung des standardisierten Distanzscorings in beiden Studien zu kaum einer Änderung

des Zusammenhangs (vgl. Tabellen 5.7 und 6.12). Auffällig war hingegen, dass sich unter Verwendung des dichotomen Scorings der Zusammenhang bei Studie 2 von ursprünglich .09 (vgl. Tabelle C4, Anhang C) auf .22 (vgl. Tabelle E8, Anhang E) erhöhte. In Studie 3 kam es hingegen auch beim dichotomen Scoring zu keiner Änderung des Zusammenhangs.

Insgesamt betrachtet zeigten sich über alle geschätzten Modelle hinweg ein mittlerer latenter Zusammenhang zwischen PU und schlussfolgerndem Denken sowie ein niedriger bis mittlerer latenter Zusammenhang zwischen PU und Merkfähigkeit, jeweils in erwarteter Richtung.

6.4 Diskussion

In der dritten Studie ergaben sich heterogene Ergebnisse im Hinblick auf die Konstruktvalidität der AcquA-PU Aufgaben. Einerseits zeigte sich ein nicht signifikanter Zusammenhang zwischen PU und SU in unerwarteter Richtung. Zudem erwies sich die gewählte Bestimmung der Akkuratheit von Persönlichkeitsbeurteilungen der MTSI-3 Zielpersonen als ungeeignet zur validen Erfassung einer Fähigkeit der Beurteiler:innen. Dies resultierte ebenfalls in nicht erwartungskonformen Zusammenhängen zu PU. Andererseits konnte durch die erneute Untersuchung des Zusammenhangs zwischen PU und schlussfolgerndem Denken sowie die zusätzliche Untersuchung des Zusammenhangs zwischen PU und Merkfähigkeit wichtige Evidenz für Konstruktvalidität der AcquA-PU Aufgaben gesammelt werden. Darüber hinaus wurde ein erster Versuch der gezielten Manipulation der Schwierigkeit zweier AcquA-PU Aufgaben vorgenommen. Auch hier ergaben sich gemischte Ergebnisse. Während die Implementierung von probabilistischen Kontingenzen nicht erfolgreich war, zeigten sich bei den Kontingenzketten vorläufige Hinweise, dass sich diese Kontingenzart zur Steigerung der Aufgabenschwierigkeit eignen könnte.

Die Itemselektion sowie weiterführende Analysen haben unter Verwendung beider Scoring-Methoden klar aufgezeigt, dass die probabilistischen Kontingenzen der Aufgabe PU1p von den Testpersonen nicht als solche erkannt oder verstanden wurden. Besonders deutlich wurde dies beim standardisierten Distanzscoring. Hier zeigte ein aus den vier Items mit korrekter Antwort $6.\overline{6}$ gebildetes Parcel im Messmodell für PU eine negative Ladung. Eine zusätzliche Betrachtung der Verteilung der Testpersonenantworten (d.h. der Rohwerte) bei diesen vier Items liefert eine post hoc Erklärung der negativen Ladung: So wählten 80 bis 86 % der Testpersonen auf der Rating-Skala von 0 (*unwahrscheinlich*) bis 10 (*wahrscheinlich*) die Werte 9 oder 10 – Antworten unter 6 wurden hingegen kaum gegeben. Unter der Annahme,

dass die fähigeren Testpersonen tendenziell die höchsten Werte ausgewählt haben, was bei den vier Items auf Grund des Distanzscorings zu höheren und damit schlechteren Itemscores führt, ist die negative Ladung des Parcels erklärbar. Es ist insgesamt anzunehmen, dass nur die zweimal im Laufe der Aneignungsphase präsentierte wahrscheinliche Reaktion aus der relevanten Kontingenz der Zielperson für die Einschätzung in der Anwendungsphase genutzt und die unwahrscheinliche Reaktion ignoriert wurde. Diese Annahme passt auch damit zusammen, dass beim dichotomen Scoring die genannten vier Items auf Grund von extremer Leichtigkeit eliminiert wurden. So ist mit dem RAM von Funder (1995) und den Annahmen über gute Zielpersonen von Human und Biesanz (2013) zu vereinbaren, dass eine zweimalige Präsentation relevanter Hinweise korrekte Schlussfolgerungen über die Zielperson erleichtert. Als Ursache für dieses insgesamt unerwartete Ergebnis kommt zum einen eine ungeeignete Konstruktion der probabilistischen Kontingenzen in Frage. Es ist denkbar, dass die Modellierung über die relative Häufigkeit der Reaktionen von den Testpersonen nicht intuitiv verstanden wurde. Andererseits kann auch eine unvollständige Instruktion eine Ursache gewesen sein. Die Testpersonen wurden in der Instruktion zu den AcquA-PU Aufgaben nicht darauf hingewiesen, dass gegebenenfalls beachtet werden muss, ob sich die Zielperson immer oder nur manchmal auf die kennengelernte Art und Weise verhält. Beide möglichen Ursachen sollten in Zukunft weiter untersucht werden, da probabilistische Kontingenzen aus theoretischer Sicht (Human & Biesanz, 2013; Schulze & Roberts, 2015) weiterhin eine vielversprechende Möglichkeit darstellen, die Aufgabenschwierigkeit gezielt zu erhöhen.

Die Ergebnisse zu den Kontingenzketten fielen nicht eindeutig aus. Beim standardisierten Distanzscoring ergaben sich durch den Vergleich der Aufgabe PU5k mit allen anderen bisher konstruierten AcquA-PU Aufgaben erste Hinweise darauf, dass sich diese Kontingenzart zum Konstruieren schwieriger Aufgaben eignen könnte. Beim dichotomen Scoring zeigten sich diese Hinweise allerdings nicht. Einschränkend muss zudem beachtet werden, dass für Aufgabe PU5k keine angemessene Vergleichsgrundlage existierte, da Aufgabe PU5 in der vorherigen Studie keine guten psychometrischen Ergebnisse erzielte und daher verworfen wurde. Der Vergleich mit den anderen Aufgaben vermag zwar erste vorläufige Hinweise zur relativen Schwierigkeit der Aufgabe PU5k liefern, allerdings existieren eine ganze Reihe an möglichen Einflussfaktoren auf die Schwierigkeit der AcquA-Aufgaben (Schulze & Roberts, 2015), sodass beim Vergleich mit anderen Aufgaben der Einfluss der Kontingenzketten nicht isoliert betrachtet werden kann. Hierzu wäre beispielsweise eine experimentelle Untersuchung notwendig, bei der nur die relevante Kontingenz variiert und die

restlichen Charakteristika der Aufgabe konstant gehalten werden. Eine ähnliche Untersuchung wäre auch für die Absicherung des Effekts von probabilistischen Kontingenzen notwendig.

Wie bereits in der vorherigen Studie wurde mit 50 % (standardisiertes Distanzscoring) und fast 60 % (dichotomes Scoring) ein sehr großer Anteil der AcquA-PU Items eliminiert. Trotzdem fiel die Reliabilität der AcquA-PU Aufgaben beim standardisierten Distanzscoring zufriedenstellend aus. Beim dichotomen Scoring war die Schätzung allerdings auffallend geringer. Ein möglicher Grund hierfür ist Aufgabe PU2, deren zweite Anwendungsphase beim dichotomen Scoring vollständig eliminiert wurde und die im Messmodell für PU eine geringe standardisierte Ladung von lediglich .26 zeigte. Trotzdem konnte bei beiden Scoring-Methoden im Vergleich zur vorherigen Studie eine Steigerung der Reliabilität erzielt werden.

Entgegen der Erwartungen deutet der geschätzte latente Zusammenhang zwischen PU und SU darauf hin, dass eine höhere Ausprägung bei PU tendenziell mit einer geringeren Ausprägung bei SU einhergeht. Dieses Ergebnis ist nicht mit der großen konzeptuellen Überlappung beider Konstrukte zu vereinbaren. Aber auch unabhängig von den konkret betrachteten Fähigkeiten stellt dies ein ungewöhnliches Ergebnis dar, da intellektuelle Fähigkeiten in der Regel positiv miteinander korrelieren (Carroll, 1993; Horn & Cattell, 1966; Schneider & McGrew, 2018). Hinzu kommt, dass sowohl PU als auch SU erwartungskonforme Zusammenhänge zu schlussfolgerndem Denken und Merkfähigkeit in derselben Richtung zeigten. Die Zusammenhänge zwischen SU und den beiden Konstrukten fielen zwar größer aus als in bisherigen Studien zum MTSI (Baumgarten, 2015; Conzelmann et al., 2015), lassen dadurch aber noch stärker vermuten, dass bei PU und SU ähnliche kognitive Prozesse eine Rolle spielen. Bei der Interpretation des Zusammenhangs zwischen PU und SU ist allerdings zu berücksichtigen, dass dieser unter Verwendung beider Scoring-Methoden nicht signifikant ausfiel. Hierbei sind natürlich auch die geringe Reliabilität der SU-Aufgaben zu bedenken sowie beim dichotomen Scoring die ebenfalls geringe Reliabilität der PU-Aufgaben. Einzeln betrachtet zeigten die SU-Aufgaben FB und HR ähnlich hohe Werte für Cronbachs α wie bei Baumgarten (2015) und zudem gute Ladungen im Faktormodell. Reliabilität und Ladung der Aufgabe RF fiel hingegen relativ gering aus. Ein Vergleich der Reliabilität von RF mit den Ergebnissen von Baumgarten (2015) war nicht möglich, da hier die Aufgabe nicht eingesetzt wurde. Bei Conzelmann et al. (2013) wird eine Reliabilität berichtet, allerdings auf Ebene der Itemscores und nicht auf Ebene der Situationsscores, die in der aktuellen Studie zur Bestimmung der Reliabilität genutzt wurden. Eine nachträgliche Bestimmung der Reliabilität von RF auf Itemebene ergab mit $\alpha = .54$, 95% CI [.43, .62] einen geringeren Wert als bei Conzelmann et al. (2013; vgl. Tabelle 6.2). Die Ergebnisse deuten somit darauf hin, dass vor

allem zu wenige SU-Aufgaben eingesetzt wurden, was auf Grund der sehr umfangreichen Studie und der langen Dauer der einzelnen SU-Aufgaben (ca. 20 bis 25 Minuten, Süß et al., 2009) nicht anders realisierbar war.

Ein positiver Zusammenhang zwischen den AcquA-PU Aufgaben und den SU-Aufgaben des MTSI-3 wäre auf Grund der konzeptuellen Überschneidungen beider Konstrukte und des realitätsnahen Materials der SU-Aufgaben (vgl. Conzelmann et al., 2013; Süß et al., 2009) ein wichtiger Hinweis auf konvergente Validität der AcquA-PU Aufgaben gewesen. Ungeklärt bleibt allerdings, ob das Ergebnis der aktuellen Studie gegen die Konstruktvalidität der PU-Aufgaben, der SU-Aufgaben oder gegebenenfalls sogar beider Instrumente spricht. In dem Zusammenhang muss berücksichtigt werden, dass – wie bereits in Abschnitt 6.2.2.2 geschildert – nur eingeschränkte Evidenz für Konstruktvalidität des MTSI-3 und seiner Vorgängerversionen vorliegt (Baumgarten, 2015; Conzelmann et al., 2013; Seidel, 2007; Weis, 2008). Hinzu kommt, dass in früheren Studien zwischen den SU-Aufgaben und Operationalisierungen konzeptuell verwandter Konstrukte bereits einzelne Zusammenhänge in unerwarteter Richtung gefunden wurden. Bei Weis (2008) zeigten sich kleine Zusammenhänge in unerwarteter Richtung zu einem Instrument zur Erfassung von nonverbaler Sensitivität (Profile of Nonverbal Sensitivity [PONS-Test]; Rosenthal et al., 1979), der Fähigkeit die Bedeutung nonverbaler Hinweise des Gesichts, des Körpers und der Stimme korrekt zu identifizieren (Hall, 2001). Die Autorin wies allerdings darauf hin, dass unter anderem die Reliabilität des PONS-Tests gering ausfiel und Belege für konvergente Validität des Instruments weitestgehend fehlen (Weis, 2008). Bei Baumgarten (2015) ergaben sich ebenfalls kleine Zusammenhänge in unerwarteter Richtung zu Skalen eines Instruments zur Erfassung der Emotionserkennungsfähigkeit (Multimodal Emotion Recognition Test [MERT]; Bänziger et al., 2009). Auch hier wurde von der Autorin unter anderem auf die geringen Reliabilitäten des MERT als mögliche Ursache hingewiesen (Baumgarten, 2015).

Dass die Akkuratheitswerte der Persönlichkeitsbeurteilungen der Zielpersonen RF, FB und HR keine eindimensionale Struktur aufweisen, lässt den Schluss zu, dass hierdurch keine gemeinsame zugrunde liegende Fähigkeit der Beurteiler:innen erfasst wurde. Eine solche Überprüfung der Eindimensionalität von Akkuratheitswerten verschiedener Zielpersonen wurde in der bisherigen Forschung nicht vorgenommen (z.B. Hall et al., 2018; Jaksic & Schlegel, 2020; Schmid Mast et al., 2011; Vogt & Colvin, 2003), sodass keine Aussage darüber möglich ist, ob dies ein ungewöhnliches Ergebnis darstellt oder nicht. Vogt und Colvin (2003) führten zumindest Analysen der Trennschärfen mehrerer Akkuratheitswerte von vier Zielpersonen durch. Von insgesamt 12 Werten wiesen allerdings nur zwei geringe

Trennschärfen auf und wurden folglich von den Autoren nicht weiter berücksichtigt. Ebenfalls konnte in bisherigen Studien keine Selektion der Akkuratheitswerte vorgenommen werden, da durch die dort verwendeten Methoden keine Itemscores resultierten (vgl. Back & Nestler, 2016; Biesanz, 2010). Die Itemselektion wurde in der vorliegenden Studie getrennt für die drei Zielpersonen vorgenommen und hat dazu geführt, dass die Akkuratheit auf Basis unterschiedlicher NEO-FFI Items bestimmt wurde. Denkbar ist, dass bei den drei Zielpersonen jeweils Items selektiert wurden, mit denen unterschiedliche Komponenten der Akkuratheit erfasst wurden. In Frage kommen hier die distinktive und normative Akkuratheit (Biesanz, 2010), die ebenfalls nur kleine Interkorrelationen zeigen (De Kock et al., 2015; Human & Biesanz, 2011). Eine weitere mögliche Erklärung für die fehlenden Zusammenhänge der Akkuratheitswerte ist, dass die Zielpersonen des MTSI-3 nicht für die Erfassung einer Fähigkeit der Beurteiler:innen geeignet waren. Nach den Ergebnissen von Rogers und Biesanz (2019) zeigen sich interindividuelle Unterschiede in der Akkuratheit nur bei guten Zielpersonen, das heißt bei solchen, über die ausreichend relevante Hinweise auf ihre Persönlichkeitseigenschaften vorliegen. Es ist nicht auszuschließen, dass dies auf die Zielpersonen des MTSI-3 nicht zutrifft. In diesem Fall ist laut RAM (Funder, 1995) unabhängig von der beurteilenden Person keine akkurate Persönlichkeitsbeurteilung und somit auch keine Erfassung interindividueller Unterschiede möglich (Rogers & Biesanz, 2019). Das Aufgabenmaterial der SU-Aufgaben wurde gesammelt, indem die Zielpersonen im Alltag ein bis zwei Tage mit einer Videokamera begleitet wurden (Conzelmann et al., 2013), und beinhaltet unter anderem Situationen beim Einkaufen, mit Freund:innen oder auf der Arbeit (Süß et al., 2009). Möglicherweise war das Verhalten hier mehr durch die Situation bestimmt als durch die Persönlichkeit der Zielpersonen, zumal diese wussten, dass sie gefilmt werden. Hinzu kommt, dass die NEO-FFI Selbstberichte der Zielpersonen als Kriterium genutzt wurden, um die Akkuratheit zu bestimmen. Möglicherweise waren die Zielpersonen nicht zu einer validen Selbstauskunft fähig oder haben sozial erwünscht geantwortet (Conzelmann et al., 2013), was eine generelle Problematik des Target-Scorings darstellt (Funder, 2012; Mayer & Geher, 1996; vgl. Abschnitt 3.2). Insbesondere sozial erwünschtes Antwortverhalten könnte bei den Zielpersonen ein großes Problem dargestellt haben, da ihre Selbstberichte nicht anonym waren. Den Zielpersonen muss bewusst gewesen sein, dass die Autor:innen des MTSI ihre Antworten auf die Items des NEO-FFI sehen, da die Persönlichkeitsprofile ausgewertet und bei Seidel (2007) und Weis (2008) unter Angabe eines Vornamen veröffentlicht wurden. Hinzu kommt, dass sich Zielpersonen und Autor:innen des MTSI kannten, da die Zielpersonen aus dem Familien- und Bekanntenkreis der Arbeitsgruppe rekrutiert wurden (Weis, 2008). Sozial

erwünschtes Antwortverhalten der Zielpersonen ist daher als sehr wahrscheinlich anzusehen, was wiederum eine starke Beeinträchtigung der Validität der mittels Selbst-Fremd Übereinstimmung bestimmten Akkuratheitswerte der vorliegenden Studie darstellt.

Darüber hinaus äußerten einige Testpersonen gegenüber der Versuchsleitung, dass ihnen der NEO-FFI Fremdbericht auf Grund der begrenzten Information schwerfiel. Das lässt vermuten, dass die Testpersonen zur Beurteilung der Persönlichkeit der Zielpersonen auch auf ihr allgemeines Wissen über die durchschnittliche Persönlichkeit oder auf Information über ihre eigene Persönlichkeit zurückgegriffen haben (vgl. Paunonen & Hong, 2013). Die zur Bestimmung der Akkuratheit verwendeten standardisierten Distanzwerte sollten ähnlich wie die Profilkorrelationen weiterhin die normative Komponente der Akkuratheit (Funder, 1999; Kenny & Winquist, 2001) und damit eine Wissenskomponente beinhalten. Auch die Ähnlichkeit zwischen Zielperson und Testperson könnte einen Einfluss auf die Akkuratheitswerte gehabt haben. Allerdings weisen die drei Zielpersonen ein recht ähnliches NEO-FFI Persönlichkeitsprofil auf (Süß et al., 2009). Die größte Abweichung findet sich bei der Gewissenhaftigkeit der Zielperson HR. Der Akkuratheitswert dieses Big Five-Faktors wurde im Rahmen der Itemselektion bei HR allerdings ausgeschlossen. Allgemeines Wissen über die durchschnittliche Persönlichkeit wie auch Ähnlichkeit zwischen Testperson und Zielperson sollten daher bei allen drei Zielpersonen einen ähnlichen Einfluss auf die Akkuratheitswerte gehabt haben.

Insgesamt ist die Validität der Interpretation der bestimmten Akkuratheitswerte fraglich. Die maximal kleinen und nicht signifikanten Zusammenhänge zwischen den AcquA-PU Aufgaben und den Akkuratheitswerten der drei Zielpersonen des MTSI-3 sind somit nicht eindeutig interpretierbar, auch wenn zumindest zwei der Zusammenhänge ein Vorzeichen in erwarteter Richtung aufwiesen.

Anders als in der vorherigen Studie wurde unter Verwendung beider Scoring-Methoden ein substantieller und signifikanter Zusammenhang zwischen PU und schlussfolgerndem Denken gefunden. Dieser deutet darauf hin, dass eine höhere Ausprägung bei PU tendenziell mit einer höheren Fähigkeit zum schlussfolgernden Denken einhergeht. Auch wenn der Zusammenhang mittel und nicht wie ursprünglich erwartet hoch ausfiel, wurde hierdurch ein wichtiger Hinweis auf konvergente Validität der AcquA-PU Aufgaben geliefert, was für die klassischen Operationalisierungen einer Fähigkeit der Beurteiler:innen nicht immer gelang (Colvin & Bundick, 2001; Letzring, 2008). Die Höhe des Zusammenhangs ist zudem vergleichbar mit der Höhe des Zusammenhangs zwischen EU und schlussfolgerndem Denken bei Hellwig et al. (2020) sowie höher als die bisher gefundenen Zusammenhänge zwischen

akkuraten Persönlichkeitsbeurteilungen und allgemeiner Intelligenz (vgl. De Kock et al., 2020). Insgesamt liefert das Ergebnis Unterstützung für die Annahme, dass PU die Fähigkeit zum logischen Schlussfolgern im Bereich Persönlichkeit darstellt. Das abweichende Ergebnis zu Studie 2 lässt sich zudem nicht auf die in beiden Studien unterschiedliche Zusammenstellung der AcquA-PU Aufgaben zurückführen. So sind die Unterschiede in einem ähnlichen Ausmaß auch dann noch vorhanden, wenn ausschließlich diejenigen PU-Aufgaben in den Analysen berücksichtigt werden, die in beiden Studien unverändert eingesetzt wurden. Auch ein Effekt von Unterschieden im klassischen logischen Schlussfolgern zwischen den Stichproben beider Studien kann ausgeschlossen werden, da beide Stichproben keine signifikanten Unterschiede im schlussfolgernden Denken, erfasst mit dem WIT-2, aufwiesen. Die geringe Reliabilität der AcquA-PU Aufgaben in Studie 2 oder andere Stichprobeneffekte erscheinen als Erklärung für die abweichenden Ergebnisse beider Studien daher am plausibelsten.

Dass der Zusammenhang zwischen PU und schlussfolgerndem Denken größer ausfiel als der Zusammenhang zwischen PU und Merkfähigkeit bestätigt, dass mit dem AcquA-Testdesign nicht primär Merkfähigkeit erfasst wird (Hellwig et al., 2020). Der Zusammenhang kann daher auch als wertvolle Evidenz für diskriminante Validität der AcquA-PU Aufgaben gewertet werden. Besonders aussagekräftig ist hier das Ergebnis der latenten Regression beim standardisierten Distanzscoring, das die zentralere Rolle des schlussfolgernden Denkens gegenüber der Merkfähigkeit deutlich gemacht hat. Dieses Ergebnis war beim dichotomen Scoring, bei dem auch der Zusammenhang zwischen PU und Merkfähigkeit etwas höher ausfiel, nicht so deutlich. Warum sich hier eine Diskrepanz zwischen den beiden Scoring-Methoden ergab, ist nicht klar und inhaltlich nicht sinnvoll zu erklären. Eventuell könnte dies an den bei beiden Scoring-Methoden unterschiedlichen selektierten AcquA-PU Items liegen, wobei immer noch eine Schnittmenge von 17 Items vorhanden ist. Berücksichtigt werden muss zudem die etwas geringere Reliabilität der AcquA-PU Aufgaben beim dichotomen Scoring.

In der dritten Studie konnte somit weitere Evidenz für konvergente und diskriminante Validität der AcquA-PU Aufgaben gesammelt werden. Allerdings fielen die Ergebnisse zur Konstruktvalidität insgesamt betrachtet heterogen aus, sodass diese als noch nicht hinreichend sichergestellt angesehen werden muss und weitere Untersuchungen folgen sollten. Während nun erste Hinweise darauf vorliegenden, dass mit den AcquA-PU Aufgaben eine Fähigkeit erfasst wird, bei der logisches Schlussfolgern eine wichtige Rolle spielt, fehlt es noch an Evidenz, dass dieses logische Schlussfolgern im Inhaltsbereich Persönlichkeit stattfindet. Die Ergebnisse der ersten gezielten Manipulation der Schwierigkeit der AcquA-PU Aufgaben

fielen ebenfalls heterogen aus, da nur für die Kontingenzketten erste und vorläufige Hinweise erzielt werden konnten, dass sich diese zur gezielten Konstruktion schwieriger Aufgaben eignen. Trotzdem wurde auch hierdurch ein Ausgangspunkt für weitere Untersuchungen theoretisch herleitbarer schwierigkeitsstiftender Merkmale geschaffen.

7. Allgemeine Diskussion

Im Zentrum der vorliegenden Thesis stand das Konstrukt PU, das auf Basis von Theorien und Modellen aus dem Bereich akkurater Persönlichkeitsbeurteilungen (z.B. RAM; Funder, 1995), relevanter angrenzender Forschungsbereiche (z.B. SI; Weis & Süß, 2005) sowie in Anlehnung an die kognitiven Operationen aus der klassischen Intelligenzforschung als Fähigkeit zum logischen Schlussfolgern über die Persönlichkeit anderer Personen konzeptualisiert wurde. PU stellt eine Integration der in den Theorien und Modellen im Zusammenhang mit akkuraten Persönlichkeitsbeurteilungen thematisierten Fähigkeiten dar und soll den zentralen kognitiven Prozess widerspiegeln, der solchen Beurteilungen zu Grunde liegt. Ausgehend von dieser Konzeptualisierung wurden im Rahmen von drei empirischen Studien Aufgaben zur Erfassung von PU konstruiert, die den Prinzipien des AcquA-Testdesigns von Schulze und Roberts (2015) folgen. Zudem wurde die psychometrische Qualität der Aufgaben überprüft sowie erste Validitätsevidenz gesammelt.

7.1 Zusammenfassende Bewertung der Studienergebnisse

Insgesamt betrachtet deuten die Ergebnisse der durchgeführten Studien darauf hin, dass die AcquA-PU Aufgaben eine reliable und in Teilen valide Erfassung einer Fähigkeit zum logischen Schlussfolgern im Bereich Persönlichkeit erlauben. In Studie 2 erwiesen sich zwei Aufgaben als psychometrisch ungeeignet und wurden daher nicht weiter berücksichtigt oder überarbeitet. Ansonsten konnte in allen drei Studien die Eindimensionalität der jeweils eingesetzten Aufgaben bestätigt werden. Es liegt somit studienübergreifende Evidenz für faktorielle Validität der AcquA-PU Aufgaben vor und es kann angenommen werden, dass diese eine einzige Fähigkeit erfassen. Über die Studien hinweg hat sich zudem gezeigt, dass eine reliable Erfassung von PU grundsätzlich möglich ist. Es muss jedoch in Abhängigkeit der gewählten Scoring-Methode auf eine ausreichende Aufgabenanzahl geachtet werden. Bei Verwendung des standardisierten Distanzscorings sowie von mindestens fünf Aufgaben fiel die Reliabilität zufriedenstellend aus. Im Gegensatz dazu reichten beim dichotomen Scoring sechs inhaltlich homogene Aufgaben für eine reliable Erfassung von PU nicht aus, sodass hier in Zukunft mehr Aufgaben benötigt werden. Dass das standardisierte Distanzscoring mit psychometrisch besseren Ergebnissen einhergeht, wurde zudem nicht nur bei Betrachtung der Reliabilität deutlich. Im Vergleich zum dichotomen Scoring zeigten sich beim standardisierten Distanzscoring ebenfalls höhere Ladungen der Aufgaben-Parcels im Messmodell für PU, höhere Trennschärfen der Items sowie eine geringere Anzahl an eliminierten Items. Diese

Ergebnisse sind in etwa vergleichbar mit denen, die auf Basis von AcquA-Aufgaben zur Erfassung von EU (Pisters & Schulze, 2017) sowie unter Verwendung von SJTs (De Leng et al., 2017) erzielt wurden. Bei den AcquA-PU Aufgaben zeigten sich allerdings deutlich größere Unterschiede zwischen beiden Scoring-Methoden, insbesondere bei den Schätzungen der Reliabilität und der Faktorladungen. Die Ergebnisse sprechen somit dafür, dass das standardisierte Distanzscoring bei der Auswertung der AcquA-PU Aufgaben bevorzugt verwendet werden sollte (vgl. aber auch Abschnitt 7.2). Dennoch sollte auch bei dieser Scoring-Methode eine Steigerung der Reliabilität durch Neukonstruktionen angestrebt werden. Da bisher ausschließlich Aufgaben vorliegen, deren relevante Kontingenzen den Big Five-Faktoren Gewissenhaftigkeit und Verträglichkeit zugeordnet werden können, sollten hierbei in erster Linie andere Persönlichkeitseigenschaften modelliert werden.

Die Umsetzung eines neuen Formats der Aufgabendarstellung kann als erfolgreich beurteilt werden. In der ersten Studie zeigten die 3D-Simulationen keinen substantiellen Methodeneffekt oder anderweitige Nachteile im Vergleich zum bisher verwendeten 2D-Format von Hellwig et al. (2020). Zudem erlauben die 3D-Simulationen eine realitätsnähere und ökologisch validere Erfassung von PU und sind daher zu bevorzugen. Diese Schlussfolgerung passt darüber hinaus zu den Ergebnissen von Jaksic und Schlegel (2020), die in ihrer Studie unter anderem die Akkuratheit von Persönlichkeitsbeurteilungen bei Verwendung unterschiedlicher Materialien (Bilder, Video ohne Ton, Video mit Ton) betrachtet haben. Hierbei zeigte sich, dass interindividuelle Unterschiede in der Akkuratheit bei Verwendung von Videos besser erfasst wurden als bei Verwendung von Bildern (Jaksic & Schlegel, 2020).

Bei den gefundenen Zusammenhängen zwischen PU und EU (Studie 1) sowie PU und logischem Schlussfolgern (Studie 3) handelt es sich um zentrale Evidenz für konvergente Validität der AcquA-PU Aufgaben. Hieraus kann geschlossen werden, dass die Aufgaben logisches Schlussfolgern in einem Inhaltsbereich erfordern, der große Gemeinsamkeiten mit EU aufweist, aber empirisch hiervon abgegrenzt werden kann. Dies entspricht den theoretisch angenommenen Beziehungen zwischen den drei Konstrukten. EU wurde entsprechend der Konzeptualisierung von Hellwig et al. (2020) als logisches Schlussfolgern im Bereich Emotionen, die ein wichtiger Bestandteil einiger Persönlichkeitseigenschaften sind (vgl. McCrae & John, 1992), aufgefasst und weist daher große konzeptuelle Gemeinsamkeiten mit PU auf. Hinzu kommt, dass in Studie 1 beide Konstrukte durch das AcquA-Testdesign operationalisiert wurden, sodass neben inhaltlichen Überschneidungen auch ein großer Methodeneffekt zu erwarten war. Dass die empirische Abgrenzung von PU und EU dennoch erfolgreich war, ist daher als besonders bedeutsame Evidenz für Konstruktvalidität zu werten.

Die Beziehung von PU und logischem Schlussfolgern ist darüber hinaus ein Befund, der sich bei Verwendung klassischer Methoden zur Erfassung akkurater Persönlichkeitsbeurteilungen trotz entsprechender Erwartungen nicht immer gezeigt hat (z.B. Letzring, 2008; vgl. auch De Kock et al., 2020), aber auch nicht immer erwartet wurde (Colvin & Bundick, 2001). Zwar resultierte ein Zusammenhang zwischen PU und logischem Schlussfolgern in annähernd erwartungskonformer Höhe nur in Studie 3 und nicht in Studie 2, allerdings ist das Ergebnis aus Studie 2 auf Grund der zu geringen Reliabilität der AcquA-PU Aufgaben mit einer größeren Unsicherheit verbunden. Das Ergebnis aus Studie 3 kann bei der Beurteilung der Konstruktvalidität der AcquA-PU Aufgaben daher höher gewichtet werden.

Weitere Evidenz für konvergente Validität konnte nicht aufgezeigt werden, da die Beziehungen von PU und SU sowie PU und akkuraten Persönlichkeitsbeurteilungen in Studie 3 nicht erwartungskonform ausfielen. Allerdings sprechen diese Ergebnisse nicht zwangsweise gegen die Konstruktvalidität der AcquA-PU Aufgaben. Wie bereits in Abschnitt 6.2.2.2 angemerkt wurde, liegt für die SU-Aufgaben des MTSI-3 von Süß et al. (2009) ebenfalls nur eingeschränkte Validitätsevidenz vor (z.B. Baumgarten, 2015; Conzelmann et al., 2013). Zudem wiesen die eingesetzten SU-Aufgaben eine zu geringe Reliabilität auf, sodass der negative und nicht signifikante Zusammenhang zwischen PU und SU mit Vorsicht interpretiert und nicht überbewertet werden sollte. Weitere Hinweise könnte eine erneute Untersuchung des Zusammenhangs unter Verwendung einer größeren Anzahl an MTSI-SU Aufgaben zur Steigerung der Reliabilität liefern. Interessant wäre auch eine Untersuchung, in der – analog zu Studie 1 – beide Konstrukte mit dem AcquA-Testdesign operationalisiert werden. Wie Hellwig (2016) bereits anmerkte, würde sich das AcquA-Testdesign auch zur Operationalisierung von SU eignen. Bei SU muss soziale Information korrekt verstanden und interpretiert werden (Weis & Süß, 2005). Es kann daher argumentiert werden, dass die beim AcquA-Testdesign im Zentrum stehende Aneignung von Wissen (Schulze & Roberts, 2015) bei SU ebenfalls eine wichtige Rolle spielt, da korrekte Schlussfolgerungen über die Emotionen, Gedanken, Intentionen, Motivationen und Persönlichkeitseigenschaften einer Zielperson (Weis, 2008) erst dann möglich sind, wenn zuvor relevantes individuelles Wissen über die Person erworben wurde. Eine solche Untersuchung würde natürlich eine zusätzliche Validierung der neu konstruierten AcquA-SU Aufgaben erfordern.

Im Hinblick auf die Ergebnisse zum Zusammenhang zwischen den AcquA-PU Aufgaben und der Akkuratheit von Persönlichkeitsbeurteilungen liegen Hinweise auf eine deutliche Beeinträchtigung der Validität der mittels Selbst-Fremd Übereinstimmung bestimmten Akkuratheitswerte vor. Nicht ausreichend relevante Hinweise auf die

Persönlichkeitseigenschaften der beurteilten Zielpersonen des MTSI-3 könnte eine akkurate Beurteilung und somit auch das Erfassen interindividueller Unterschiede in der Akkuratheit verhindert haben (vgl. Funder, 1995; Rogers & Biesanz, 2019). Zudem spricht die fehlende Anonymität beim Selbstbericht der Zielpersonen über ihre eigene Persönlichkeit (vgl. Weis, 2008) gegen die Validität der in Studie 3 bestimmten Akkuratheitswerte. Folglich lassen sich aus den nicht erwartungskonformen Zusammenhängen zwischen den AcquA-PU Aufgaben und den Akkuratheitswerten auch keinerlei Schlüsse über die Validität der AcquA-PU Aufgaben ziehen. Da die Bestimmung der Akkuratheit von Persönlichkeitsbeurteilungen als klassische Operationalisierung einer Fähigkeit der Beurteiler:innen aufgefasst werden kann (vgl. Biesanz, 2010; Cronbach, 1955; Funder, 1999; Hall et al., 2018), wäre ein moderater positiver Zusammenhang ein wichtiger Hinweis auf konvergente Validität der AcquA-PU Aufgaben gewesen. Sinnvoll wäre daher eine erneute Untersuchung dieses Zusammenhanges mit neuem Material über reale Zielpersonen, in dem sichergestellt wird, dass relevante Hinweise auf deren Persönlichkeit zur Verfügung stehen. Eine Möglichkeit wäre die Orientierung an der Studie von Christiansen et al. (2005) und der dort vorgeschlagenen Bestimmung der Interview Accuracy. Auf Basis von Jobinterviews konstruierten die Autor:innen einen Multiple-Choice-Test mit Fragen zu persönlichkeitsrelevanten Verhaltensweisen der interviewten Personen. Die Verwendung von Interviews als Material zur Erfassung der Fremdbeurteilungen hat den entscheidenden Vorteil, dass über gezielte Interviewfragen die Generierung relevanter Hinweise zu einem gewissen Ausmaß sichergestellt werden könnte. Die korrekten Antworten wurden bei Christiansen et al. (2005) allerdings ebenfalls über das problembehaftete Target-Scoring festgelegt. Durch die zusätzliche Verwendung von Fremdberichten (z.B. von guten Bekannten der Zielperson) könnte die Validität der Akkuratheitswerte entsprechend den Empfehlungen von Funder (1995) zur Verwendung mehrerer Kriterien allerdings erhöht werden.

Evidenz für diskriminante Validität der AcquA-PU Aufgaben lieferten die Untersuchungen der Zusammenhänge zwischen PU und den Big Five-Faktoren. Unter Verwendung zweier verschiedener Operationalisierungen der Big Five zeigten sich sowohl in Studie 1 als auch in Studie 2 lediglich kleine und überwiegend nicht signifikante Zusammenhänge zwischen PU und Verträglichkeit, Offenheit für Erfahrungen und Gewissenhaftigkeit. Die Ergebnisse sind somit in etwa vergleichbar mit den Zusammenhängen zwischen Gf und den Big Five (Ackerman & Heggestad, 1997). Die deskriptiv etwas höheren Zusammenhänge zwischen PU und Verträglichkeit sowie Offenheit passen allerdings eher zu den in anderen Studien gefundenen Zusammenhängen zwischen Verträglichkeit und der

Normative Accuracy (Biesanz, 2010; Letzring, 2015) sowie Offenheit und DI (Christiansen et al., 2005; vgl. aber auch de Vries, et al., 2021) beziehungsweise Offenheit und Gc (Ackerman & Heggestad, 1997) und somit weniger zu der angenommenen kognitiven Operation bei PU. Die Zusammenhänge könnten daher als erste vorläufige Hinweise angesehen werden, dass normative Akkuratheit beziehungsweise Wissensaspekte bei den AcquA-PU Aufgaben einen gewissen Einfluss auf die Testleistung haben könnten. Dies sind allerdings nur sehr vage Hinweise, die näher untersucht werden müssen, beispielsweise indem die Rolle von persönlichkeitsbezogenem Wissen (DI; Christiansen et al., 2005) näher betrachtet wird.

Ein Effekt der tatsächlichen Ähnlichkeit zwischen Testperson und Zielperson im Hinblick auf den Big Five-Faktor Gewissenhaftigkeit konnte weder in Studie 1 noch in Studie 2 gefunden werden. In Studie 2 ergaben sich wiederum vorläufige Hinweise darauf, dass es einen Zusammenhang zwischen dem Interesse der Testpersonen an den Emotionen und Sorgen anderer sowie der Richtigkeit von Schlussfolgerungen über bestimmte Aspekte der Verträglichkeit der Zielpersonen gibt. Ein nicht vorhandener Effekt der Ähnlichkeit zwischen Testperson und Zielperson spricht allerdings nicht gegen die Validität der AcquA-PU Aufgaben. Auch wenn Studien einen Zusammenhang zwischen der tatsächlichen oder angenommenen Ähnlichkeit und der Akkuratheit von Persönlichkeitsbeurteilungen gezeigt haben (Funder et al., 1995; Vogt & Colvin, 2003), ist dies kein konsistenter Befund (vgl. Hartung & Renner, 2011; Kurtz & Sherker, 2003). Zudem scheint ein solcher Zusammenhang vorwiegend für den normativen Aspekt der Akkuratheit vorzuliegen (Human & Biesanz, 2011, 2012; Letzring, 2015), der mit den AcquA-PU Aufgaben ohnehin nicht erfasst werden soll.

Wichtige diskriminante Validitätsevidenz lieferte darüber hinaus der Zusammenhang zwischen PU und Merkfähigkeit. In Studie 3 ergab sich ein kleiner bis mittlerer Zusammenhang, der geringer ausfiel als der Zusammenhang zwischen PU und schlussfolgerndem Denken. Das Einprägen und Erinnern typischer Verhaltensweisen der Zielperson kann zwar als eine Voraussetzung für PU angesehen werden, sollte aber nicht die primäre kognitive Anforderung der AcquA-PU Aufgaben darstellen, wenn diese PU valide erfassen sollen. Insbesondere die Ergebnisse der latenten Regression lassen den Schluss zu, dass dies nicht der Fall ist, wenngleich die Beziehung von PU und Merkfähigkeit beim dichotomen Scoring auffällig größer ausfiel als beim standardisierten Distanzscoring. Während Merkfähigkeit beziehungsweise Gedächtnis auch im Zusammenhang mit SI untersucht und in das Integrative Modell Sozialer Intelligenz aufgenommen wurde (Weis & Süß, 2005), findet eine solche Fähigkeit im Bereich akkurater Persönlichkeitsbeurteilungen bisher kaum Berücksichtigung. Zwar wird die Rolle eines guten Gedächtnisses von Autor:innen erwähnt

(Christiansen et al., 2005), allerdings existieren kaum Studien, die den Zusammenhang untersucht haben (Colman, 2021). Die vorhandene Evidenz aus Studie 3 kann somit als Vorteil der AcquA-PU Aufgaben gegenüber den bisherigen Methoden zur Bestimmung akkurater Persönlichkeitsbeurteilungen angesehen werden, bei denen die Rolle der Merkfähigkeit oder des Gedächtnisses noch vollkommen unklar ist.

Neben den insgesamt betrachtet eher optimistischen Ergebnissen zur Reliabilität und Validität der AcquA-PU Aufgaben, ergaben sich studienübergreifend auch solche, die als problematisch angesehen werden müssen und daher Einschränkungen darstellen. So musste in allen drei Studien, aber vor allem in Studie 2 und 3, ein auffällig großer Anteil der Items der AcquA-PU Aufgaben eliminiert werden. Die Gründe hierfür sind weitestgehend unklar. Aufschluss könnte eine qualitative Untersuchung des Antwortprozesses geben, in der die Testpersonen berichten, wie sie zu ihrer Antwort gelangt sind und welche Information sie hierfür verwendet haben. Zudem könnten so auch wertvolle Hinweise für Überarbeitungen gesammelt werden, die für diejenigen Items in Betracht gezogen werden sollten, die sich konsistent in mehreren Studien und unter Verwendung beider Scoring-Methoden als psychometrisch ungeeignet erwiesen haben. Ein Vergleich der Ergebnisse der Itemselektionen der Studien 2 und 3 bei denjenigen Aufgaben, die in beiden Studien unverändert eingesetzt wurden, zeigt, dass es eine größere Anzahl solcher Items gibt. Für diese würde sich eine zeitnahe Überarbeitung lohnen, um die Anzahl eliminierter Items in folgenden Studien zu reduzieren. Diese hätte allerdings auch zur Folge, dass die meisten Vertonungen der Dialoge neu erstellt werden müssten, da die Sprecher:innen größtenteils nicht mehr rekrutiert werden können. Dass nachträgliche Überarbeitungen der Aufgaben nicht ohne Weiteres durchgeführt werden können, stellt einen generellen Nachteil der AcquA-PU Aufgaben im vertonten 3D-Format dar. Dies erschwert zudem gezielte Manipulationen des Aufgabenmaterials, die zur experimentellen Untersuchung schwierigkeitsstiftender Merkmale und zur Erhebung experimenteller Validitätsevidenz in zukünftigen Studien notwendig werden könnten.

Ein zweites problematisches Ergebnis, das studienübergreifend aufgetreten ist, betrifft die Aufgabenschwierigkeit. Durch die zusätzliche Anwendung des dichotomen Scorings wurde in allen drei Studien anhand der Aufgabenmittelwerte deutlich, dass die AcquA-PU Aufgaben überwiegend extrem leicht sind. Dies war auch dann noch der Fall, wenn im Rahmen der Itemselektion extrem leichte Items mit einem Mittelwert > .95 eliminiert wurden (vgl. Studie 2 und Studie 3). Und auch wenn beim standardisierten Distanzscoring die Aufgabenmittelwerte keine ganz so eindeutige Interpretation zulassen, deuten hier die Verteilungen der Aufgaben-Parcels ebenfalls darauf hin, dass die Aufgaben eher leicht sind. Um Varianzeinschränkungen

und damit einhergehende Konsequenzen für Zusammenhangsanalysen zu vermeiden sowie zukünftig auch in den höheren Fähigkeitsbereichen zwischen Personen differenzieren zu können, ist es unerlässlich, die Aufgabenschwierigkeit weiter zu untersuchen und vor allem zu erhöhen. In Studie 3 wurde durch die Verwendung von zwei Kontingenztypen, die aus theoretischer Sicht eine Steigerung der Schwierigkeit zur Folge haben sollten, dieses Problem erstmalig, wenngleich eher unsystematisch, angegangen. Hier zeigten sich erste Hinweise darauf, dass sich Kontingenzketten zur Konstruktion schwieriger Aufgaben eignen könnten. Die probabilistischen Kontingenzen gingen hingegen nicht mit einer höheren Aufgabenschwierigkeit einher, was aber vermutlich weniger an dem Kontingenztyp an sich lag. Die Ergebnisse lassen vielmehr den Schluss zu, dass die gewählte Umsetzung ungeeignet war und die Testpersonen die probabilistische Art der Kontingenzen nicht verstanden haben.

Darüber hinaus müssen bei der Interpretation der Ergebnisse hinsichtlich der externen Validität Einschränkungen durch die Stichprobenzusammensetzung beachtet werden. In allen drei Studien waren Frauen und Studierende deutlich überrepräsentiert. Dies führte auch dazu, dass überwiegend junge Testpersonen mit guter Schulbildung untersucht wurden. Insbesondere in Studie 3 bestand die Stichprobe fast ausschließlich aus Studierenden unterschiedlicher Fachrichtungen. Hinzu kommt bei Studie 3, dass die Stichprobengröße auf Basis der vorab durchgeführten Poweranalysen als nicht ausreichend bewertet werden muss. Eine Fortführung der Erhebungen war auf Grund von zeitlichen Grenzen und der ohnehin durch die Pandemiesituation bedingten sehr langen Studiendauer nicht möglich.

7.2 Erfassung von PU mit dem AcquA-Testdesign: Vorteile und Limitationen

Die ersten Ergebnisse zur psychometrischen Qualität der neu konstruierten AcquA-PU Aufgaben sowie die erste Validitätsevidenz sprechen dafür, dass in der vorliegenden Arbeit eine gute Grundlage für eine reliable und valide Erfassung von logischem Schlussfolgern im Bereich Persönlichkeit geschaffen wurde. Durch die Konstruktion entsprechend den Prinzipien des AcquA-Testdesigns von Schulze und Roberts (2015) weisen die AcquA-PU Aufgaben wesentliche Vorteile gegenüber den in der bisherigen Forschung verwendeten Methoden zur Operationalisierung einer entsprechenden Fähigkeit auf. Das Zwei-Phasen-Design der Aufgaben ermöglicht eine eindeutige Bewertung der Testpersonenantworten (Hellwig et al., 2020; Schulze & Roberts, 2015) und somit die Erfassung von PU durch einen klassischen Leistungstest. Zur Bestimmung der Akkuratheit von Persönlichkeitsbeurteilungen wurden bisher unabhängig vom konkreten Ansatz in der Regel Selbstberichte und Fremdberichte der

Persönlichkeit miteinander verglichen (Funder, 2012). Allerdings kann angezweifelt werden, dass der Vergleich zweier Fragebögen zur Erfassung von typischem Verhalten einen geeigneten Ansatz zur Erfassung von maximalem Verhalten darstellt. Hinzu kommt, dass sowohl die Selbstberichte als auch die Fremdberichte verzerrt sein können (Funder, 2012; Kenny, 1994; Mayer & Geher, 1996), sodass deren Vergleich keine eindeutige Bewertung der Testpersonenantworten zulässt. Mit den AcquA-PU Aufgaben wird hingegen eine passende Operationalisierung der diskutierten Fähigkeit zur Verfügung gestellt.

Die Bewertung der Testpersonenantworten erfolgt bei den AcquA-PU Aufgaben ausschließlich auf Basis der in der Aneignungsphase präsentierten Information (Schulze & Roberts, 2015), sodass keinerlei Abhängigkeiten von bestimmten Persönlichkeitstheorien oder -modellen bestehen. Bei den bisher verwendeten Methoden werden die Items des für den Selbst- und Fremdbericht verwendeten Instruments vor der Bestimmung der Akkuratheit oftmals erst zu einem Skalenwert je Persönlichkeitseigenschaft zusammengefasst (z.B. Hall et al., 2016, 2018; Jaksic & Schlegel, 2020; Schmid Mast et al., 2011; siehe aber auch Funder et al., 1995; Letzring, 2008). Hier hängt die Bestimmung der Akkuratheit somit auch von dem Modell ab, das der Skalenbildung zugrunde gelegt wurde. Die AcquA-PU Aufgaben sind durch die Konstruktionsweise hingegen mit verschiedenen Theorien und Modellen zu vereinbaren. Dabei ist es zudem unerheblich, ob das persönlichkeitsrelevante Verhalten als transsituativ stabil angenommen wird oder die Theorie explizit von Variabilität des Verhaltens über verschiedene Situationen hinweg ausgeht (z.B. Kammrath et al., 2005; Mischel & Shoda, 1995; vgl. auch Hellwig et al., 2020). In der Anwendungsphase wird ein Ereignis konstruiert, das mit dem Ereignis der relevanten Kontingenz aus der Aneignungsphase vergleichbar ist (Schulze & Roberts, 2015), sodass situative Aspekte stabil gehalten werden. Dass die AcquA-PU Aufgaben mit verschiedenen Ansätzen zu vereinbaren sind, liegt zudem daran, dass die Persönlichkeitseigenschaften der Zielpersonen auf Ebene konkreter Verhaltensweisen modelliert werden. Im Gegensatz dazu wurden die Persönlichkeitsbeurteilungen in der bisherigen Forschung zum Teil durch Verwendung von extrem kurzen Verfahren mit ein bis zwei Items pro Persönlichkeitseigenschaft und somit auf einer sehr globalen und abstrakten Ebene erhoben (z.B. Hall et al., 2016, 2018; Süß et al., 2009). Dieses Vorgehen hat nicht nur aus methodischer Sicht Nachteile. Insbesondere wenn die Beurteilungen durch Items erhoben werden, die lediglich die Labels der einzelnen Persönlichkeitseigenschaften beinhalten, erfordert die akkurate Beurteilung nicht nur individuelles Wissen über die Zielperson und die Fähigkeit, korrekte Schlussfolgerungen zu ziehen. Zusätzlich wird allgemeines Wissen darüber benötigt, welche Verhaltensweisen indikativ für welche Persönlichkeitseigenschaft sind (vgl.

DI; Christiansen et al., 2005), was wiederum modellabhängig sein kann (z.B. FFM vs. HEXACO-Modell; Ashton & Lee, 2007; McCrae & John, 1992). Bei den AcquA-PU Aufgaben ist ein solches allgemeines Wissen für die korrekte Bearbeitung nicht erforderlich.

Neben diesen bedeutsamen Vorteilen der AcquA-PU Aufgaben sollen auch potentielle Einschränkungen nicht unerwähnt bleiben. Die korrekten Antworten einer Aufgabe werden auf Basis der Information aus der Aneignungsphase festgelegt und stellen konkrete Werte auf der Rating-Skala von 0 (*unwahrscheinlich*) bis 10 (*wahrscheinlich*) dar. Bei deterministischen Kontingenzen sind dies stets die Werte 0 oder 10, bei probabilistischen Kontingenzen kann dies auch ein Wert zwischen den Extrema sein (Schulze & Roberts, 2015). Um die Festlegung der korrekten Antwort eindeutig vornehmen zu können, wird die Vergleichbarkeit des für die Kontingenz relevanten Ereignisses in beiden Phasen einer Aufgabe vorausgesetzt. Die Vergleichbarkeit ist eine Annahme, die bisher nicht weiter überprüft, sondern auf Basis einer sorgfältigen Aufgabenkonstruktion getroffen wurde. Da die Ereignisse beider Phasen allerdings nie identisch sind, da ansonsten Merkfähigkeit erfasst werden würde, könnte die Frage entstehen, ob die Festlegung eines spezifischen Wertes auf der Rating-Skala gerechtfertigt ist. Situative Aspekte können Einfluss auf das (persönlichkeitsrelevante) Verhalten einer Person haben (Gilbert et al., 1988; Mischel & Shoda, 1995; Trope, 1986). Beispielsweise wird dem CAPS von Mischel und Shoda (1995) zufolge das Persönlichkeitssystem durch verschiedene Merkmale einer Situation aktiviert und erzeugt so letztendlich spezifische Verhaltensweisen. Die Autoren gehen davon aus, dass selbst bei denselben äußeren Gegebenheiten einer Situation das Aktivierungsmuster des Persönlichkeitssystems niemals identisch ist (Mischel & Shoda, 1995). Folglich muss man diesem Ansatz nach davon ausgehen, dass jede kleine Veränderung in einer Situation Einfluss auf das Verhalten hat und somit auch auf die Festlegung der korrekten Antwort auf der Rating-Skala der AcquA-PU Aufgaben. Diese potentielle Einschränkung ist allerdings keine Einschränkung des AcquA-Testdesigns an sich, sondern vielmehr der verwendeten Scoring-Methode. Die Festlegung eines konkreten Werts auf der Rating-Skala benötigen nur Methoden, die auf Distanzwerten basieren. Beim dichotomen Scoring muss hingegen ausschließlich festgelegt werden, ob die Reaktion der Zielperson in der Anwendungsphase eher wahrscheinlich (Wert > 5 auf der Rating-Skala) oder eher unwahrscheinlich ist (Wert < 5 auf der Rating-Skala; Hellwig et al., 2020). Das dichotome Scoring ist somit auch dann zu rechtfertigen, wenn man davon ausgehen muss, dass Unterschiede im Ereignis zwischen Aneignungs- und Anwendungsphase Einfluss auf das Verhalten der Zielperson haben. Während das standardisierte Distanzscoring also aus psychometrischer Sicht zu bevorzugen ist

(vgl. Abschnitt 7.1), hat das dichotome Scoring aus theoretischer und konzeptueller Sicht Vorteile. Abgesehen von den in Abschnitt 7.1 genannten Ergebnissen, unter anderem im Zusammenhang mit der Reliabilität der AcquA-PU Aufgaben, haben in der vorliegenden Arbeit beide Scoring-Methoden zu ähnlichen Ergebnissen geführt. Ob die wenigen Abweichungen zwischen den Methoden (z.B. bei der Itemselektion oder dem Zusammenhang zwischen PU und Merkfähigkeit) auf die mögliche konzeptuelle Einschränkung des Distanzscorings oder eher die geringere Reliabilität der Aufgaben beim dichotomen Scoring zurückzuführen sind, bedarf weiterer Forschung. Die Sicherstellung der Vergleichbarkeit zwischen Aneignungs- und Anwendungsphase ist in jedem Fall eine der größten Herausforderungen bei der Aufgabenkonstruktion. Eine mögliche Verbesserung für zukünftige Konstruktionen könnte darin bestehen, die Vergleichbarkeit beider Phasen systematischer sicherzustellen, indem vorab zentrale Merkmale des relevanten Ereignisses identifiziert und zwischen den Phasen konstant gehalten werden. Hierbei könnte eine Orientierung an Situations-Taxonomien wie dem Riverside Situational Q-Sort (RSQ; Funder, 2016) in Betracht gezogen werden. Das RSQ wurde mit dem Ziel entwickelt, Situationen zu beschreiben und das Ausmaß der Ähnlichkeit zwischen zwei Situationen bezüglich verschiedener psychologisch bedeutsamer Eigenschaften einschätzen zu können (Funder, 2016; Sherman et al., 2010). Die Items des RSQ könnten beispielsweise als Ausgangspunkt verwendet werden, um die zentralen und verhaltensrelevanten Merkmale eines Ereignisses zu identifizieren. Ähnliches gilt für die DIAMONDS-Taxonomie von Rauthmann et al. (2014), die aufbauend auf dem RSQ acht Dimensionen situativer Eigenschaften annimmt und für die ebenfalls Erhebungsinstrumente entwickelt wurden (Rauthmann et al., 2014; Rauthmann & Sherman, 2016).

Eine weitere zentrale Annahme, die bei den AcquA-PU Aufgaben getroffen werden muss, betrifft ebenfalls die Rolle der Situation. Für eine valide Darstellung der Persönlichkeitseigenschaften der Zielpersonen ist es erforderlich, dass die Testpersonen das Verhalten auch als persönlichkeitsrelevantes und typisches Verhalten der Zielpersonen wahrnehmen und nicht als rein durch die Situation bedingt (vgl. RAM; Funder, 1995; Letzring & Funder, 2021). Bei der Verwendung realer Zielpersonen besteht allerdings eine noch größere Einschränkung als bei den fiktiven Zielpersonen der AcquA-Aufgaben. Erfolgt die Persönlichkeitsbeurteilung auf Basis von Videomaterial, das die Zielperson in Interaktion mit anderen Personen (z.B. Funder et al., 1995; Letzring, 2008) oder im Rahmen von Interviews zeigt (z.B. De Kock et al., 2015; Hall et al., 2015), ist nicht auszuschließen, dass die Situation größeren Einfluss auf das Verhalten ausgeübt hat als die Persönlichkeit. Hinzu kommt, dass solche Situationen oftmals künstlich sind und den Zielpersonen bewusst ist, dass sie gefilmt

werden. Daher sollte bei realen Zielpersonen auch der potentielle Einfluss sozialer Erwünschtheit nicht außer Acht gelassen werden. Bei realen Zielpersonen kann man nie mit Sicherheit feststellen, ob und welchen Einfluss situative Aspekte oder soziale Erwünschtheit auf das Verhalten haben, da man hierfür auf die problembehafteten Selbst- oder Fremdberichte zurückgreifen müsste. Bei den AcquA-PU Aufgaben werden hingegen fiktive Zielpersonen verwendet, deren Persönlichkeit flexibel spezifiziert werden kann. Somit kann festgelegt werden, dass es sich um persönlichkeitsrelevantes Verhalten handelt und diese Information wiederum in die Aufgaben integriert werden. In den AcquA-PU Aufgaben der vorliegenden Arbeit wurde den Testpersonen beispielsweise explizit mitgeteilt, dass diese davon ausgehen können, dass das typische Verhalten der Zielpersonen beobachtet werden konnte.

Insgesamt betrachtet weisen die AcquA-PU Aufgaben, vor allem durch die Möglichkeit zur eindeutigen Bewertung der Testpersonenantworten, klare Vorteile gegenüber den in der bisherigen Forschung verwendeten Methoden auf. Die Vorteile gehen mit zu treffenden Annahmen einher, die potentielle Einschränkungen darstellen. Allerdings können diese durch eine sorgfältige und systematische Aufgabenkonstruktion größtenteils kontrolliert werden.

7.3 Zukünftige Forschung

Zukünftige Forschungsaktivitäten rund um die AcquA-PU Aufgaben sollten in erster Linie in den folgenden drei Bereichen stattfinden: Aufgabenkonstruktion, Steigerung der Aufgabenschwierigkeit und Sammlung weiterer Validitätsevidenz. Wie bereits geschildert, ist es erforderlich, den Aufgabenpool zu erweitern und neue Aufgaben zu konstruieren. Diese sollten zum einen andere Persönlichkeitseigenschaften in den Kontingenzen widerspiegeln und zum anderen eine höhere Schwierigkeit aufweisen. Neben weiteren Erprobungen der Kontingenzketten und einer überarbeiteten Konstruktion probabilistischer Kontingenzen, sollten auch weitere schwierigkeitsstiftende Elemente untersucht werden. Aus theoretischer Sicht käme die Festlegung von mindestens zwei relevanten Kontingenzen einer Zielperson in Frage, die in der Anwendungsphase für eine korrekte Schlussfolgerung miteinander kombiniert werden müssen. Beim klassischen logischen Schlussfolgern geht eine Steigerung der zu verarbeitenden Informationsmenge mit erhöhten Itemschwierigkeiten einher (Freund et al., 2008; Primi, 2001; Holzman et al., 1983; Loe et al., 2018; vgl. Abschnitt 6.1). Durch die Steigerung der Anzahl der relevanten Kontingenzen könnte bei den AcquA-PU Aufgaben ebenfalls die Informationsmenge gesteigert werden. Zudem ist dies mit der Annahme von Letzring und Funder (2021) zu vereinbaren, dass im letzten Schritt des RAM Hinweise

miteinander kombiniert werden müssen, wozu das Kombinieren von Hinweisen aus mindestens zwei Kontingenzen gezählt werden kann. Letzring und Funder (2021) nehmen zudem an, dass im letzten Schritt des Persönlichkeitsbeurteilungsprozesses auch andere mögliche Ursachen für das Verhalten, wie situative Einflüsse, in Betracht gezogen werden müssen. Denkbar wäre zur Steigerung der Aufgabenschwierigkeit somit auch das Modellieren situativer Einflüsse auf das Verhalten der Zielperson, die in der Anwendungsphase neben den Kontingenzen mitberücksichtigt werden müssen. Auch hier würde die zu verarbeitende Informationsmenge und somit vermutlich auch die Aufgabenschwierigkeit gesteigert. Klare situative Einflüsse ließen sich beispielsweise durch die Konstruktion von sogenannten starken Situationen modellieren (Snyder & Ickes, 1985). Starke Situationen verlangen von allen Personen ein und dasselbe Verhalten und minimieren somit Einflüsse der Persönlichkeit auf das Verhalten (Snyder & Ickes, 1985).

Bezüglich der Konstruktvalidität der AcquA-PU Aufgaben wurde bereits die erneute Untersuchung des Zusammenhangs zur Akkuratheit von Persönlichkeitsbeurteilungen als wichtige Evidenz für konvergente Validität genannt. Daneben wäre auch die Untersuchung des Zusammenhangs zwischen PU und persönlichkeitsrelevantem Wissen, beispielsweise in Form der DI von Christiansen et al. (2005; De Kock et al., 2015; de Vries et al., 2021), von Interesse. PU und DI unterscheiden sich zwar in der angenommenen zentralen kognitiven Operation (logisches Schlussfolgern vs. Wissen), teilen aber denselben Inhaltsbereich, sodass ein positiver Zusammenhang zu erwarten wäre. Dieser sollte zudem höher ausfallen als der Zusammenhang von PU und Wissen in anderen Inhaltsbereichen. Die Höhe des Zusammenhangs zwischen PU und DI könnte zudem Hinweise darauf liefern, ob es sich hierbei, wie bisher angenommen, um zwei Konstrukte handelt, die zwar Überschneidungen aufweisen, aber theoretisch voneinander getrennt werden können, oder ob diese eher zwei Komponenten einer einzelnen Fähigkeit darstellen. Letzteres wurde von Hellwig et al. (2020) für EU vorgeschlagen. Die Autor:innen nehmen für EU das logische Schlussfolgern im Bereich Emotionen als zentrale Komponente an, die durch eine Wissenskomponente bezüglich typischer emotionaler Reaktionen ergänzt wird. Im Hinblick auf PU könnte man argumentieren, dass hierzu auch die Fähigkeit gehört, zu erkennen, wann man auf allgemeines Wissen über Persönlichkeit zurückgreifen kann (z.B., weil die Zielperson ein durchschnittliches Profil aufweist; Vogt & Colvin, 2003) oder muss (z.B., weil Information fehlt; Paunonen & Hong, 2013). Im Gegensatz dazu wird im Bereich SI soziales Wissen ebenfalls getrennt von SU konzeptualisiert (Weis & Süß, 2005). Allerdings handelt es sich bei dem im Integrativen Modell zur Sozialen Intelligenz angenommenen Wissen um prozedurales

Wissen (Weis & Süß, 2005) und nicht wie bei EU und DI um explizites Wissen (Christiansen et al., 2005; Hellwig et al., 2020). Des Weiteren ist für zukünftige Untersuchungen der Überschneidungsbereich zwischen PU und EU von Interesse. Um diesen näher zu beleuchten, könnten AcquA-Aufgaben konstruiert werden, in denen die relevanten Kontingenzen der Zielpersonen Persönlichkeitseigenschaften widerspiegeln, bei denen Emotionen eine zentrale Rolle spielen (z.B. Neurotizismus; McCrae & John, 1992). Sollten die in Abschnitt 3.1.1 formulierten Annahmen über den Überschneidungsbereich beider Konstrukte korrekt sein, so müsste eine solche Aufgabe im CFA-Modell ähnlich hohe Ladungen sowohl auf PU als auch EU zeigen. Darüber hinaus wurden die AcquA-EU und AcquA-PU Aufgaben in Studie 1 in zwei separaten Blöcken mit separaten Instruktionen präsentiert, die noch einmal auf die Hauptcharakteristik der Aufgaben aufmerksam gemacht haben. Ein weiterer kritischer Test für die AcquA-PU Aufgaben sowie das AcquA-Testdesign wäre daher eine Untersuchung, in der beide Aufgabentypen durchmischt und mit einer gemeinsamen Instruktion präsentiert werden.

Sobald weitere Evidenz für die Konstruktvalidität der AcquA-PU Aufgaben vorliegt, sollte auch die Kriteriumsvalidität näher betrachtet werden. Ein kritischer Aspekt ist hierbei die Auswahl einiges geeigneten Kriteriums. Von Interesse könnte der Zusammenhang zur Beziehungsqualität von Ehepartner:innen sein, die bereits konsistente Zusammenhänge zur Selbst-Fremd Übereinstimmung gezeigt hat (Letzring & Noftle, 2010). Einige Autor:innen verwendeten zudem die Akkuratheit von Persönlichkeitsbeurteilungen als Kriterium (z.B. Christiansen et al., 2005). Folglich kann der Zusammenhang zwischen PU und der Akkuratheit auch als eine Form der Kriteriumsvalidität interpretiert werden, je nachdem, ob die Akkuratheit als klassische Operationalisierung einer Fähigkeit aufgefasst wird oder nicht.

Letztendlich sollte auch experimentelle Validitätsevidenz gesammelt werden, was bisher nicht möglich war. Neben experimentellen Untersuchungen schwierigkeitsstiftender Merkmale könnte eine erneute Orientierung an den Moderatoren des RAM (Funder, 1995) ein guter Anhaltspunkt sein. Der Moderator Good Information bezieht sich unter anderem auf die Quantität der für die Beurteiler:innen verfügbaren Information und deren Auswirkung auf die Akkuratheit (Funder, 1995). Mit steigender Quantität nimmt in der Regel die Akkuratheit zu, was auch als *Acquaintanceship Effect* bezeichnet wurde (Funder et al., 1995; Krzyzaniak et al., 2019; Letzring et al., 2006). Demzufolge ist bei den AcquA-PU Aufgaben eine bessere Leistung zu erwarten, wenn die Kontingenzen in der Aneignungsphase nicht nur einmal, sondern mehrfach präsentiert werden. Hierfür ist allerdings zunächst einmal eine grundsätzliche Steigerung der Aufgabenschwierigkeit erforderlich, damit ein solcher Effekt überhaupt in den Daten sichtbar werden kann.

8. Zusammenfassung und Fazit

Die Frage, welche Personen gute Beurteiler:innen der Persönlichkeit anderer Personen sind, ist eine der ersten und zentralsten, der in der Forschung rund um die Akkuratheit von Persönlichkeitsbeurteilungen nachgegangen wurde (Funder, 2012). Und auch eine Fähigkeit, akkurate Beurteilungen vornehmen zu können, war bereits früh von Interesse (z.B. Taft, 1955; Vernon, 1933). Trotzdem existiert bisher keine einheitliche theoretische Vorstellung davon, wie eine solche Fähigkeit beschrieben werden kann. Die vorhandenen Beschreibungen sind eher unpräzise (z.B. Funder, 1995), weisen konzeptuelle Schwierigkeiten auf (z.B. Mayer, 2008, 2009) oder stimmen nicht hinsichtlich der angenommenen zentralen kognitiven Operation überein (z.B. Christiansen et al., 2005; Weis & Süß, 2005). Gleichzeitig wurden sehr heterogene Ansätze vorgeschlagen, um die Akkuratheit von Persönlichkeitsbeurteilungen zu erfassen (vgl. Back & Nestler, 2016; Biesanz, 2010; Hall et al., 2018), die darüber hinaus alle von einem grundlegenden Problem betroffen sind: Unabhängig vom spezifischen Ansatz steht kein Kriterium zur Verfügung, das eine eindeutige Bestimmung der wahren Persönlichkeit der Zielperson und damit der Akkuratheit zulässt (z.B. Funder, 2012). Folglich eignet sich keiner der in der bisherigen Forschung vorgeschlagenen Ansätze zur Operationalisierung einer Fähigkeit der Beurteiler:innen. Ausgehend von diesen zwei Problematiken wurde in der vorliegenden Arbeit zunächst auf Basis vorhandener Modelle und Theorien (u.a. Funder, 1995; Kammrath et al., 2005; Weis & Süß, 2005) das Konstrukt PU abgeleitet und als zentrale Fähigkeit vorgeschlagen, die akkuraten Persönlichkeitsbeurteilungen zugrunde liegt. PU wurde hierbei als Fähigkeit zum logischen Schlussfolgern über die Persönlichkeit anderer Personen aufgefasst. Durch eine Konzeptualisierung unter Berücksichtigung der kognitiven Operationen aus dem klassischen Intelligenzbereich (vgl. Schneider & McGrew, 2018) wurde eine fundierte theoretische Grundlage geschaffen, um begründete Vorhersagen über Zusammenhänge zu anderen Fähigkeiten sowie insbesondere eine dem Konstrukt angemessene Operationalisierung abzuleiten. Durch die Anwendung des AcquA-Testdesigns von Schulze und Roberts (2015) bei der Aufgabenkonstruktion konnte zudem die seit langem im Bereich akkurater Persönlichkeitsbeurteilungen bestehende und kritisch diskutierte Kriteriumsproblematik (z.B. Bernieri, 2001; Funder, 1999; Kenny, 1994; Vernon, 1933) angegangen und gelöst werden, was die Relevanz der Arbeit für den Forschungsbereich deutlich macht. AcquA-Aufgaben zur Erfassung von PU bestehen aus zwei Phasen: In der ersten Phase wird den Testpersonen präsentiert, mit welchen typischen Verhaltensweisen eine Zielperson auf bestimmte Ereignisse reagiert. Diese Ereignis-Verhaltens-Kontingenzen werden verwendet, um vorab festgelegte

Persönlichkeitseigenschaften der Zielperson zu modellieren. Im Anschluss wird die Zielperson in einer neuen, aber mit der ersten Phase vergleichbaren Situation präsentiert und die Testpersonen müssen für mehrere mögliche Reaktionen der Zielperson einschätzen, ob diese wahrscheinlich oder unwahrscheinlich wären. Auf Basis der Information aus der ersten Phase ist es möglich, die Testpersonenantworten eindeutig und logikbasiert als richtig oder falsch zu bewerten (Schulze & Roberts, 2015; siehe auch Hellwig et al., 2020).

Im Rahmen von drei empirischen Studien wurden AcquA-Aufgaben zur Erfassung von PU konstruiert, überarbeitet, die psychometrische Qualität untersucht und erste korrelative Validitätsevidenz gesammelt. In Studie 1 (N = 202) wurden zunächst drei AcquA-PU Aufgaben in einer leicht überarbeiteten Version des 2D-Formats von Hellwig et al. (2020) sowie drei Aufgaben in einem neuen 3D-Format konstruiert. Zudem wurde der Zusammenhang zwischen PU und EU sowie PU und der Persönlichkeit der Testpersonen untersucht. Es ergaben sich Hinweise auf Eindimensionalität und damit faktorielle Validität der neu konstruierten Aufgaben. Unter Verwendung des standardisierten Distanzscorings resultierte zudem eine zufriedenstellende Reliabilität. Darüber hinaus zeigte sich ein erwartungskonform großer latenter Zusammenhang zwischen AcquA-Aufgaben zur Erfassung von PU und EU. Letztere wurden ebenfalls sowohl im 2D-Format als auch im 3D-Format eingesetzt, sodass der potentielle Effekt des neues Aufgabenformats untersucht werden konnte. Hierzu wurde ein CTC(M-1)-Modell (Eid, 2000; Eid et al., 2003) geschätzt, in dem sich kein systematischer beziehungsweise substantieller Einfluss eines Methodenfaktors zeigte. Darüber hinaus resultierten auf latenter Ebene maximal geringe und damit weitestgehend erwartungskonforme Zusammenhänge zwischen den AcquA-PU Aufgaben und Skalen der Big Five-Faktoren. Es zeigte sich zudem kein höherer Zusammenhang zwischen PU und Persönlichkeit der Testperson, wenn Zielperson und Testperson eine ähnliche Ausprägung auf Facetten der Gewissenhaftigkeit aufweisen.

In Studie 2 (N = 204) wurden die vorhandenen Aufgaben im 3D-Format vertont und um drei neue Aufgaben ergänzt. Zudem wurde der Zusammenhang von PU zu klassischem logischen Schlussfolgern sowie erneut zur Persönlichkeit der Testpersonen untersucht. Zwei der neu konstruierten Aufgaben zeigten keine guten psychometrischen Ergebnisse und wurden daher nicht weiter berücksichtigt. Für die übrigen Aufgaben konnte erneut Evidenz für faktorielle Validität gesammelt werden, wohingegen die Reliabilität unter Verwendung beider Scoring-Methoden nicht zufriedenstellend ausfiel. Darüber hinaus konnte der erwartete große latente Zusammenhang zwischen den AcquA-PU Aufgaben und einem Maß für klassisches logisches Schlussfolgern nicht aufgezeigt werden. Im Gegensatz dazu zeigten sich erneut

maximal geringe latente Zusammenhänge zu den Big Five. Des Weiteren ergaben sich erste vorläufige Hinweise auf einen möglichen Effekt der tatsächlichen Ähnlichkeit zwischen Testperson und Zielperson im Hinblick auf spezifische Facetten der Verträglichkeit.

In Studie 3 (N = 129 bzw. 162) wurde erneut der Zusammenhang zwischen PU und klassischem logischen Schlussfolgern sowie zusätzlich der Zusammenhang zwischen PU und Merkfähigkeit untersucht. Zudem stand die Beziehung von PU und SU sowie PU und akkuraten Persönlichkeitsbeurteilungen im Fokus. Auf latenter Ebene zeigte sich eine mittlere Korrelation zwischen PU und klassischem logischem Schlussfolgern, die erwartungsgemäß höher ausfiel als der Zusammenhang zwischen PU und Merkfähigkeit. Entgegen der Erwartungen zeigte sich ein nicht signifikanter Zusammenhang von PU und SU in unerwarteter Richtung. Darüber hinaus resultierten Ergebnisse, die auf mangelnde Validität der bestimmten Akkuratheitswerte hindeuteten. Dies ging mit nicht erwartungskonformen Zusammenhängen zwischen den AcquA-PU Aufgaben und der Akkuratheit von Persönlichkeitsbeurteilungen einher. In Studie 3 wurde zudem ein erster Versuch der gezielten Manipulation der Schwierigkeit zweier AcquA-PU Aufgaben vorgenommen. Während die Konstruktion probabilistischer Kontingenzen nicht erfolgreich war, zeigten sich bei den verwendeten Kontingenzketten vorläufige Hinweise, dass sich diese Kontingenzart zur Steigerung der Aufgabenschwierigkeit eignen könnte. Darüber hinaus resultierte in der dritten Studie unter Verwendung des standardisierten Distanzscorings erneut eine zufriedenstellende Reliabilität der fünf verwendeten Aufgaben.

Zusammenfassend betrachtet deuten die Ergebnisse der drei durchgeführten Studien darauf hin, dass die konstruierten AcquA-Aufgaben zur Erfassung von PU eine geeignete Grundlage darstellen, auf der man zukünftig aufbauen kann, um eine reliable und valide Erfassung einer Fähigkeit zum logischen Schlussfolgern über die Persönlichkeit anderer Personen zu ermöglichen.

9. Literaturverzeichnis

Ackerman, P. L. & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, *121*(2), 219–245. https://doi.org/10.1037/0033-2909.121.2.219

Adams, H. F. (1927). The good judge of personality. *The Journal of Abnormal and Social Psychology*, *22*(2), 172–181. https://doi.org/10.1037/h0075237

Allport, G. W. (1961). *Pattern and growth in personality*. Holt, Rinehart and Winston.

Allport, G. W. (1985). The historical background of social psychology. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (3rd ed., Vol. 1, pp. 1–46). Random House.

Ambady, N., Hallahan, M. & Rosenthal, R. (1995). On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology*, *69*(3), 518–529. https://doi.org/10.1037/0022-3514.69.3.518

Asendorpf, J. B., Banse, R. & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology*, *83*(2), 380–393. https://doi.org/10.1037/0022-3514.83.2.380

Asendorpf, J. B. & Ostendorf, F. (1998). Is self-enhancement healthy? Conceptual, psychometric, and empirical analysis. *Journal of Personality and Social Psychology*, *74*(4), 955–966. https://doi.org/10.1037/0022-3514.74.4.955

Ashton, M. C. & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*(2), 150–166. https://doi.org/10.1177/1088868306294907

Ashton, M. C. & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, *91*(4), 340–345. https://doi.org/10.1080/00223890902935878

Ashton, M. C., Lee, K. & de Vries, R. E. (2014). The HEXACO honesty-humility, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, *18*(2), 139–152. https://doi.org/10.1177/1088868314523838

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M. & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin & Review*, *22*(2), 366–377. https://doi.org/10.3758/s13423-014-0699-x

Back, M. D. & Nestler, S. (2016). Accuracy of judging personality. In J. A. Hall, M. Schmid Mast & T. V. West (Eds.), *The social psychology of perceiving others accurately* (pp. 98–124). Cambridge University Press. https://doi.org/10.1017/CBO9781316181959.005

Back, M. D., Schmukle, S. C. & Egloff, B. (2008). How extraverted is honey.bunny77@hotmail.de? Inferring personality from e-mail addresses. *Journal of Research in Personality*, *42*(4), 1116–1122. https://doi.org/10.1016/j.jrp.2008.02.001

Bagby, R. M., Rector, N. A., Bindseil, K., Dickens, S. E., Levitan, R. D. & Kennedy, S. H. (1998). Self-report ratings and informants' ratings of personalities of depressed outpatients. *American Journal of Psychiatry*, *155*(3), 437–438. https://doi.org/10.1176/ajp.155.3.437

Banse, R. & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636. https://doi.org/10.1037/0022-3514.70.3.614

Bänziger, T., Grandjean, D. & Scherer, K. R. (2009). Emotion recognition from expressions in face, voice, and body: The Multimodal Emotion Recognition Test (MERT). *Emotion*, *9*(5), 691–704. https://doi.org/10.1037/a0017088

Barrick, M. R., Patton, G. K. & Haugland, S. N. (2000). Accuracy of interviewer judgments of job applicant personality traits. *Personnel Psychology*, *53*(4), 925–951. https://doi.org/10.1111/j.1744-6570.2000.tb02424.x

Baumgarten, M. (2015). *Soziales Verständnis—Theoretische Grundlagen, Konstruktion und Validierung zweier Testaufgaben zum Kernkonstrukt der sozialen Intelligenz.* punkt um FILM.

Beauducel, A. & Kersting, M. (2002). Fluid and crystallized intelligence and the Berlin Model of Intelligence Structure (BIS). *European Journal of Psychological Assessment*, *18*(2), 97–112. https://doi.org/10.1027/1015-5759.18.2.97

Beer, A. & Watson, D. (2008). Personality judgment at zero acquaintance: Agreement, assumed similarity, and implicit simplicity. *Journal of Personality Assessment*, *90*(3), 250–260. https://doi.org/10.1080/00223890701884970

Bernieri, F. J. (2001). Toward a taxonomy of interpersonal sensitivity. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 3–20). Lawrence Erlbaum Associates Publishers.

Biesanz, J. C. (2010). The Social Accuracy Model of interpersonal perception: Assessing individual differences in perceptive and expressive accuracy. *Multivariate Behavioral Research*, *45*(5), 853–885. https://doi.org/10.1080/00273171.2010.519262

Blackman, M. C. (2002). Personality judgment and the utility of the unstructured employment interview. *Basic and Applied Social Psychology*, *24*(3), 241–250. https://doi.org/10.1207/153248302760179156

Blanch-Hartigan, D. & Hill Cummings, K. (2021). Training and improving accuracy of personality trait judgments. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford handbook of accurate personality judgment* (pp. 307–318). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190912529.013.21

Bleidorn, W. & Ostendorf, F. (2009). Ein Big Five-Inventar für Kinder und Jugendliche: Die deutsche Version des Hierarchical Personality Inventory for Children (HiPIC). *Diagnostica*, *55*(3), 160–173. https://doi.org/10.1026/0012-1924.55.3.160

Block, J. (1961). *The Q-sort method in personality assessment and psychiatric research*. Charles C Thomas Publisher. https://doi.org/10.1037/13141-000

Borkenau, P. & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology*, *62*(4), 645–657. https://doi.org/10.1037/0022-3514.62.4.645

Borkenau, P., Mosch, A., Tandler, N. & Wolf, A. (2016). Accuracy of judgments of personality based on textual information on major life domains. *Journal of Personality*, *84*(2), 214–224. https://doi.org/10.1111/jopy.12153

Borkenau, P. & Ostendorf, F. (2008). *NEO-Fünf-Faktoren-Inventar nach Costa und McCrae (NEO-FFI): Manual* (2. Aufl.). Hogrefe.

Brandt, N. D., Becker, M., Tetzner, J., Brunner, M., & Kuhl, P. (2021). What teachers and parents can add to personality ratings of children: Unique associations with academic performance in elementary school. *European Journal of Personality*, *35*(6), 814-832. https://doi.org/10.1177/0890207020988436

Browne, M. W. & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). University of California Press.

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, *109*, Article 104047. https://doi.org/10.1016/j.jml.2019.104047

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Carver, R. P. (1990). *Reading rate: A review of research and theory.* Academic Press.

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading*, *36*(2), 84–95.

Cattell, R. B. (1979). *Personality and learning theory: Vol. 1. The structure of personality in its environment*. Springer Publishing Company.

Chen, R., Rafaeli, E., Bar-Kalifa, E., Gilboa-Schechtman, E., Lutz, W. & Atzil-Slonim, D. (2018). Moderators of congruent alliance between therapists and clients: A realistic accuracy model. *Journal of Counseling Psychology*, *65*(6), 703–714. https://doi.org/10.1037/cou0000285

Christiansen, N. D., Wolcott-Burnam, S., Janovics, J. E., Burns, G. N. & Quirk, S. W. (2005). The good judge revisited: Individual differences in the accuracy of personality judgments. *Human Performance*, *18*(2), 123–149. https://doi.org/10.1207/s15327043hup1802_2

Colman, D. E. (2021). Characteristics of the judge that are related to accuracy. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford handbook of accurate personality judgment.* (pp. 85–99). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190912529.013.6

Colvin, C. R. & Bundick, M. J. (2001). In search of the good judge of personality: Some methodological and theoretical concerns. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 47–65). Lawrence Erlbaum Associates Publishers.

Connelly, B. S. & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, *136*(6), 1092–1122. https://doi.org/10.1037/a0021212

Conzelmann, K., Weis, S. & Süß, H.-M. (2013). New findings about social intelligence: Development and application of the Magdeburg Test of Social Intelligence (MTSI). *Journal of Individual Differences*, *34*(3), 119–137. https://doi.org/10.1027/1614-0001/a000106

Costa, P. T., Jr. & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences*, *13*(6), 653–665. https://doi.org/10.1016/0191-8869(92)90236-I

Cronbach, L. J. (1955). Processes affecting scores on „understanding of others“ and „assumed similarity“. *Psychological Bulletin*, *52*(3), 177–193. https://doi.org/10.1037/h0044919

Cronbach, L. J. (1958). Proposals leading to analytic treatment of social perception scores. In R. Tagiuri & L. Petrullo (Eds.), *Person perception and interpersonal behavior* (pp. 353–379). Stanford University Press.

Cronbach, L. J. (1992). Four Psychological Bulletin articles in perspective. *Psychological Bulletin*, *112*(3), 389–392. https://doi.org/10.1037/0033-2909.112.3.389

Cronbach, L. J. & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin*, *50*(6), 456–473. https://doi.org/10.1037/h0057173

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16–29. https://doi.org/10.1037/1082-989X.1.1.16

Darbyshire, D., Kirk, C., Wall, H. J. & Kaye, L. K. (2016). Don’t judge a (Face)book by its cover: Exploring judgement accuracy of others’ personality on Facebook. *Computers in Human Behavior, 58*, 380–387. https://doi.org/10.1016/j.chb.2016.01.021

Davis, M. H. & Kraus, L. A. (1997). Personality and empathic accuracy. In W. J. Ickes (Ed.), *Empathic accuracy* (pp. 144–168). The Guilford Press.

Deal, J. E., Halverson, C. F., Martin, R. P., Victor, J. & Baker, S. (2007). The Inventory of Children’s Individual Differences: Development and validation of a short version. *Journal of Personality Assessment*, *89*(2), 162–166. https://doi.org/10.1080/00223890701468550

De Kock, F. S., Lievens, F. & Born, M. P. (2015). An in-depth look at dispositional reasoning and interviewer accuracy. *Human Performance*, *28*(3), 199–221. https://doi.org/10.1080/08959285.2015.1021046

De Kock, F. S., Lievens, F. & Born, M. P. (2017). A closer look at the measurement of dispositional reasoning: Dimensionality and invariance across assessor groups. *International Journal of Selection and Assessment*, *25*(3), 240–252. https://doi.org/10.1111/ijsa.12176

De Kock, F. S., Lievens, F. & Born, M. P. (2020). The profile of the ‘good judge’ in HRM: A systematic review and agenda for future research. *Human Resource Management Review*, *30*(2), Article 100667. https://doi.org/10.1016/j.hrmr.2018.09.003

De Leng, W. E., Stegers-Jager, K. M., Husbands, A., Dowell, J. S., Born, M. P. & Themmen, A. P. N. (2017). Scoring method of a situational judgment test: Influence on internal consistency reliability, adverse impact and correlation with personality? *Advances in Health Sciences Education*, *22*(2), 243–265. https://doi.org/10.1007/s10459-016-9720-7

de Vries, R. E., Barends, A. J. & de Kock, F. S. (2021). Dispositional insight: Its relations with HEXACO personality and cognitive ability. *Personality and Individual Differences*, *173*, Article 110644. https://doi.org/10.1016/j.paid.2021.110644

DeYoung, C. G., Quilty, L. C. & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*(5), 880–896. https://doi.org/10.1037/0022-3514.93.5.880

Digman, J. M. (1990). Personality structure: Emergence of the Five-Factor Model. *Annual Review of Psychology*, *41*(1), 417–440. https://doi.org/10.1146/annurev.ps.41.020190.002221

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, *65*(2), 241–261. https://doi.org/10.1007/BF02294377

Eid, M., Lischetzke, T. & Nussbeck, F. W. (2006). Structural equation models for multitrait-multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology*. (pp. 283–299). American Psychological Association. https://doi.org/10.1037/11383-020

Eid, M., Lischetzke, T., Nussbeck, F. W. & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods*, *8*(1), 38–60. https://doi.org/10.1037/1082-989X.8.1.38

Evans, T. R., Hughes, D. J. & Steptoe-Warren, G. (2020). A conceptual replication of emotional intelligence as a second-stratum factor of intelligence. *Emotion*, *20*(3), 507–512. http://dx.doi.org/10.1037/emo0000569

Eysenck, H. J. (1992). Four ways five factors are not basic. *Personality and Individual Differences*, *13*(6), 667–673. https://doi.org/10.1016/0191-8869(92)90237-J

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Fleeson, W. & Noftle, E. (2008). The end of the person-situation debate: An emerging synthesis in the answer to the consistency question. *Social and Personality Psychology Compass*, *2*(4), 1667–1684. https://doi.org/10.1111/j.1751-9004.2008.00122.x

Freund, P. A., Hofer, S. & Holling, H. (2008). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological Measurement*, *32*(3), 195–210. https://doi.org/10.1177/0146621607306972

Friesen, C. A. & Kammrath, L. K. (2011). What it pays to know about a close other: The value of if-then personality knowledge in close relationships. *Psychological Science*, *22*(5), 567–571. https://doi.org/10.1177/0956797611405676

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*(4), 652–670. https://doi.org/10.1037/0033-295x.102.4.652

Funder, D. C. (1999). *Personality Judgment. A realistic approach to person perception*. Academic Press.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, *21*(3), 177–182. https://doi.org/10.1177/0963721412445309

Funder, D. C. (2016). Taking situations seriously: The situation construal model and the Riverside situational Q-sort. *Current Directions in Psychological Science*, *25*(3), 203–208. https://doi.org/10.1177/0963721416635552

Funder, D. C., Kolar, D. C. & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology*, *69*(4), 656–672. https://doi.org/10.1037/0022-3514.69.4.656

Funder, D. C. & West, S. G. (1993). Consensus, self-other agreement, and accuracy in personality judgment: An introduction. *Journal of Personality*, *61*(4), 457–476. https://doi.org/10.1111/j.1467-6494.1993.tb00778.x

Furr, R. M. (2008). A framework for profile similarity: Integrating similarity, normativeness, and distinctiveness. *Journal of Personality*, *76*(5), 1267–1316. https://doi.org/10.1111/j.1467-6494.2008.00521.x

Gage, N. L. & Cronbach, L. J. (1955). Conceptual and methodological problems in interpersonal perception. *Psychological Review*, *62*(6), 411–422. https://doi.org/10.1037/h0047205

Gardner, H. (1983). Frames of mind: A theory of multiple intelligences. Basic Books.

Gardner, H. (1991). *Abschied vom IQ. Die Rahmen-Theorie der vielfachen Intelligenzen*. Klett-Cotta.

Gardner, H. & Hatch, T. (1989). Multiple intelligences go to school: Educational implications of the theory of multiple intelligences. *Educational Researcher*, *18*(8), 4-10. https://doi.org/10.2307/1176460

Gardner, H., & Moran, S. (2006). The science of multiple intelligences theory: A response to Lynn Waterhouse. *Educational Psychologist, 41*(4), 227–232. https://doi.org/10.1207/s15326985ep4104_2

Gilbert, D. T. & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*(1), 21–38. https://doi.org/10.1037/0033-2909.117.1.21

Gilbert, D. T., Pelham, B. W. & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology, 54*(5), 733–740. https://doi.org/10.1037/0022-3514.54.5.733

Goldberg, L. R. (n.d.). *IPIP scale-construction procedures*. International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences. https://ipip.ori.org/newScaleConstruction.htm

Goldberg, L. R. (1990). An alternative „description of personality“: The Big-Five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1219–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*(1), 26–34. https://doi.org/10.1037/0003-066X.48.1.26

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. J. Deary, F. De Fruyt & F. Ostendorf (Eds.), *Personality Psychology in Europe* (Vol. 7, pp. 7–28). Tilburg University Press.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R. & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*(1), 84–96. https://doi.org/10.1016/j.jrp.2005.08.007

Goldberg, L. R. & Velicer, W. F. (2006). Principles of exploratory factor analysis. In S. Strack (Ed.), *Differentiating normal and abnormal personality* (2nd ed., pp. 209–237). Springer Publishing Company.

Gorsuch, R. L. (2008). *Factor analysis* (2nd ed.). Psychology Press.

Gosling, S. D., Ko, S. J., Mannarelli, T. & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology, 82*(3), 379–398. https://doi.org/10.1037/0022-3514.82.3.379

Göttert, R. & Asendorpf, J. (1989). Eine deutsche Version des California-Child-Q-Sort Kurzform. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *21*(1), 70–82.

Graziano, W. G., & Tobin, R. M. (2018). Agreeableness: A three-level integration. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *The SAGE handbook of personality and individual differences: Vol. 3. Applications of personality and individual differences* (pp. 212–234). SAGE Reference. https://doi.org/10.4135/9781526451248.n9

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

Guilford, J. P. (1985). The structure-of-intellect model. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurement, and applications* (pp. 225–266). Wiley.

Guilford, J. P. (1988). Some changes in the structure-of-intellect model. *Educational and Psychological Measurement*, *48*(1), 1–4. https://doi.org/10.1177/001316448804800102

Hall, J. A. (2001). The PONS Test and the psychometric approach to measuring interpersonal sensitivity. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 143–160). Lawrence Erlbaum Associates Publishers.

Hall, J. A., Back, M. D., Nestler, S., Frauendorfer, D., Schmid Mast, M. & Ruben, M. A. (2018). How do different ways of measuring individual differences in zero-acquaintance personality judgment accuracy correlate with each other? *Journal of Personality*, *86*(2), 220–232. https://doi.org/10.1111/jopy.12307

Hall, J. A., Goh, J. X., Schmid Mast, M. & Hagedorn, C. (2016). Individual differences in accurately judging personality from text. *Journal of Personality*, *84*(4), 433–445. https://doi.org/10.1111/jopy.12170

Hall, J. A., Gunnery, S. D., Letzring, T. D., Carney, D. R. & Colvin, C. R. (2017). Accuracy of judging affect and accuracy of judging personality: How and when are they related? *Journal of Personality*, *85*(5), 583–592. https://doi.org/10.1111/jopy.12262

Halverson, C. F., Havill, V. L., Deal, J., Baker, S. R., Victor, J. B., Pavlopoulos, V., Besevegis, E. & Wen, L. (2003). Personality structure as derived from parental ratings of free descriptions of children: The Inventory of Child Individual Differences. *Journal of Personality*, *71*(6), 995–1026. https://doi.org/10.1111/1467-6494.7106005

Hammond, K. R. & Stewart, T. R. (Eds.). (2001). *The essential Brunswik: Beginnings, explications, applications*. Oxford University Press.

Hartig, J., Jude, N. & Rauch, W. (2003). *Entwicklung und Erprobung eines deutschen Big-Five-Fragebogens auf Basis des International Personality Item Pools (IPIP40)* (Arbeiten aus dem Institut für Psychologie der Johann Wolfgang Goethe-Universität, Heft 1, 2003). Institut für Psychologie der Johann Wolfgang Goethe-Universität.

Hartung, F.-M. & Renner, B. (2011). Social curiosity and interpersonal perception: A judge × trait interaction. *Personality and Social Psychology Bulletin*, *37*(6), 796–814. https://doi.org/10.1177/0146167211400618

Hartwig, M. & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*(4), 643–659. https://doi.org/10.1037/a0023589

Hellwig, S. (2016). *Die Erfassung von Emotional Understanding mit dem Empathic Agent Paradigma* [Dissertation, Bergische Universität Wuppertal]. Hochschulschriften der Bergischen Universität Wuppertal. http://elpub.bib.uni-wuppertal.de/edocs/dokumente/fbg/psychologie/diss2016/hellwig/dg1601.pdf

Hellwig, S., Roberts, R. D. & Schulze, R. (2020). A new approach to assessing emotional understanding. *Psychological Assessment*, *32*(7), 649–662. https://doi.org/10.1037/pas0000822

Hellwig, S. & Schulze, R. (2021). Emotion theories as a scoring rationale for tests of emotional understanding. *Personality and Individual Differences*, *181*, Article 111034. https://doi.org/10.1016/j.paid.2021.111034

Hirschmüller, S., Egloff, B., Nestler, S. & Back, M. D. (2013). The dual lens model: A comprehensive framework for understanding self–other agreement of personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, *104*(2), 335–353. https://doi.org/10.1037/a0030383

Hofstee, W. K. B., de Raad, B. & Goldberg, L. R. (1992). Integration of the Big Five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, *63*(1), 146–163. https://doi.org/10.1037/0022-3514.63.1.146

Holzman, T. G., Pellegrino, J. W. & Glaser, R. (1983). Cognitive variables in series completion. *Journal of Educational Psychology*, *75*(4), 603–618. https://doi.org/10.1037/0022-0663.75.4.603

Horn, J. L. & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*(5), 253–270. https://doi.org/10.1037/h0023816

Horn, J. L. & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53–91). The Guilford Press.

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

Huffcutt, A. I., Conway, J. M., Roth, P. L. & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, *86*(5), 897–913. https://doi.org/10.1037/0021-9010.86.5.897

Human, L. J. & Biesanz, J. C. (2011). Through the looking glass clearly: Accuracy and assumed similarity in well-adjusted individuals' first impressions. *Journal of Personality and Social Psychology*, *100*(2), 349–364. https://doi.org/10.1037/a0021850

Human, L. J. & Biesanz, J. C. (2012). Accuracy and assumed similarity in first impressions of personality: Differing associations at different levels of analysis. *Journal of Research in Personality*, *46*(1), 106–110. https://doi.org/10.1016/j.jrp.2011.10.002

Human, L. J. & Biesanz, J. C. (2013). Targeting the good target: An integrative review of the characteristics and consequences of being accurately perceived. *Personality and Social Psychology Review*, *17*(3), 248–272. https://doi.org/10.1177/1088868313495593

Human, L. J., Biesanz, J. C., Finseth, S. M., Pierce, B. & Le, M. (2014). To thine own self be true: Psychological adjustment promotes judgeability via personality–behavior congruence. *Journal of Personality and Social Psychology*, *106*(2), 286–303. https://doi.org/10.1037/a0034860

Human, L. J., Mignault, M.-C., Biesanz, J. C. & Rogers, K. H. (2019). Why are well-adjusted people seen more accurately? The role of personality-behavior congruence in naturalistic social settings. *Journal of Personality and Social Psychology*, *117*(2), 465–482. https://doi.org/10.1037/pspp0000193

Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen: Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells. *Diagnostica*, *28*(3), 195–225.

Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychologische Rundschau*, *35*(1), 21–35.

Jäger, A. O., Holling, H., Preckel, F., Schulze, R., Vock, M., Süß, H.-M. & Beauducel, A. (2006). *Berliner Intelligenzstruktur-Test für Jugendliche: Begabungs- und Hochbegabungsdiagnostik (BIS-HB): Manual*. Hogrefe.

Jäger, A. O., Süß, H.-M. & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test Form 4 (BIS-4): Handanweisung.* Hogrefe.

Jaksic, C. & Schlegel, K. (2020). Accuracy in judging others' personalities: The role of emotion recognition, emotion understanding, and trait emotional intelligence. *Journal of Intelligence*, *8*(3), Article 34. https://doi.org/10.3390/jintelligence8030034

John, O. P., Angleitner, A. & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality*, *2*(3), 171–203. https://doi.org/10.1002/per.2410020302

John, O. P., Naumann, L. P. & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114–158). The Guilford Press.

John, O. P. & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). The Guilford Press.

Johnson-Laird, P. N. (1994). A model theory of induction. *International Studies in the Philosophy of Science*, *8*(1), 5–29. https://doi.org/10.1080/02698599408573474

Kammrath, L. K., Mendoza-Denton, R. & Mischel, W. (2005). Incorporating if...then...personality signatures in person perception: Beyond the person-situation dichotomy. *Journal of Personality and Social Psychology*, *88*(4), 605–618. https://doi.org/10.1037/0022-3514.88.4.605

Kang, S.-M., Day, J. D. & Meara, N. M. (2005). Social and emotional intelligence: Starting a conversation about their similarities and differences. In R. Schulze & R. D. Roberts (Eds.), *Emotional intelligence: An international handbook* (pp. 91–105). Hogrefe & Huber Publishers.

Karelaia, N. & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404–426. https://doi.org/10.1037/0033-2909.134.3.404

Kaufmann, E. & Athanasou, J. A. (2009). A meta-analysis of judgment achievement as defined by the lens model equation. *Swiss Journal of Psychology*, *68*(2), 99–112. https://doi.org/10.1024/1421-0185.68.2.99

Kaufmann, E., Reips, U.-D. & Wittmann, W. W. (2013). A critical meta-analysis of lens model studies in human judgment and decision-making. *PLoS ONE*, *8*(12), Article e83528. https://doi.org/10.1371/journal.pone.0083528

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods*, *39*(4), 979–984. https://doi.org/10.3758/BF03192993

Kelley, K. (2020). *The MBESS R package* (R package version 4.7.0 and higher) [Computer software]. https://cran.r-project.org/package=MBESS

Kelley, K. & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods*, *21*(1), 69–92. https://doi.org/10.1037/a0040086

Kenny, D. A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review*, *98*(2), 155–163. https://doi.org/10.1037/0033-295x.98.2.155

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. The Guilford Press.

Kenny, D. A. (2004). PERSON: A general model of interpersonal perception. *Personality and Social Psychology Review*, *8*(3), 265–280. https://doi.org/10.1207/s15327957pspr0803_3

Kenny, D. A. & Winquist, L. (2001). The measurement of interpersonal sensitivity: Consideration of design, components, and unit of analysis. In J. A. Hall & F. J. Bernieri (Eds.), *Interpersonal sensitivity: Theory and measurement* (pp. 265–302). Lawrence Erlbaum Associates Publishers.

Kersting, M., Althoff, K. & Jäger, A. O. (2008). *Wilde-Intelligenz-Test 2 (WIT-2): Manual*. Hogrefe.

Kim, H., Di Domenico, S. I. & Connelly, B. S. (2019). Self–other agreement in personality reports: A meta-analytic comparison of self- and informant-report means. *Psychological Science*, *30*(1), 129–138. https://doi.org/10.1177/0956797618810000

Klein, P. D. (1997). Multiplying the problems of intelligence by eight: A critique of Gardner's theory. *Canadian Journal of Education*, *22*(4), 377–394. https://doi.org/10.2307/1585790

Klein, P. D. (1998). A response to Howard Gardner: Falsifiability, empirical evidence, and pedagogical usefulness in educational psychologies. *Canadian Journal of Education*, *23*(1), 103–112. https://doi.org/10.2307/1585969

Krzyzaniak, S. L., Colman, D. E., Letzring, T. D., McDonald, J. S. & Biesanz, J. C. (2019). The effect of information quantity on distinctive accuracy and normativity of personality trait judgments. *European Journal of Personality*, *33*(2), 197–213. https://doi.org/10.1002/per.2196

Küfner, A. C. P., Back, M. D., Nestler, S. & Egloff, B. (2010). Tell me a story and I will tell you who you are! Lens model analyses of personality and creative writing. *Journal of Research in Personality*, *44*(4), 427–435. https://doi.org/10.1016/j.jrp.2010.05.003

Kurtz, J. E. & Sherker, J. L. (2003). Relationship quality, trait similarity, and self-other agreement on personality ratings in college roommates. *Journal of Personality*, *71*(1), 21–48. https://doi.org/10.1111/1467-6494.t01-1-00005

Laukka, P., Elfenbein, H. A., Thingujam, N. S., Rockstuhl, T., Iraki, F. K., Chui, W. & Althoff, J. (2016). The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features. *Journal of Personality and Social Psychology*, *111*(5), 686–705. https://doi.org/10.1037/pspi0000066

Lee, G., Dunbar, S. B. & Frisbie, D. A. (2001). The relative appropriateness of eight measurement models for analyzing scores from tests composed of testlets. *Educational and Psychological Measurement*, *61*(6), 958–975. https://doi.org/10.1177/00131640121971590

Lee, K. & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment*, *25*(5), 543–556. https://doi.org/10.1177/1073191116659134

Lee, K., Ashton, M. C., Pozzebon, J. A., Visser, B. A., Bourdage, J. S. & Ogunfowora, B. (2009). Similarity and assumed similarity in personality reports of well-acquainted persons. *Journal of Personality and Social Psychology*, *96*(2), 460–472. https://doi.org/10.1037/a0014059

Legree, P. J. (1995). Evidence for an oblique social intelligence factor established with a Likert-based testing procedure. *Intelligence*, *21*(3), 247–266. https://doi.org/10.1016/0160-2896(95)90016-0

Legree, P. J., Kilcullen, R., Psotka, J., Putka, D. & Ginter, R. N. (2010). *Scoring situational judgment tests using profile similarity metrics* (Technical Report Nr. 1272). United States Army Research Institute for the Behavioral and Social Sciences.

Legree, P. J., Psotka, J., Tremble, T. & Bourne, D. R. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze & R. D. Roberts (Eds.), *Emotional intelligence: An international handbook* (pp. 155–179). Hogrefe & Huber Publishers.

Letzring, T. D. (2008). The good judge of personality: Characteristics, behaviors, and observer accuracy. *Journal of Research in Personality*, *42*(4), 914–932. https://doi.org/10.1016/j.jrp.2007.12.003

Letzring, T. D. (2015). Observer judgmental accuracy of personality: Benefits related to being a good (normative) judge. *Journal of Research in Personality*, *54*, 51–60. https://doi.org/10.1016/j.jrp.2014.05.001

Letzring, T. D. & Funder, D. C. (2018). Interpersonal accuracy in trait judgments. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *The SAGE handbook of personality and individual differences: Vol. 3. Applications of personality and individual differences* (pp. 253–282). SAGE Reference. http://dx.doi.org/10.4135/9781526451248.n11

Letzring, T. D. & Funder, D. C. (2021). The realistic accuracy model. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford handbook of accurate personality judgment* (pp. 9–22). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190912529.013.2

Letzring, T. D. & Noftle, E. E. (2010). Predicting relationship quality from self-verification of broad personality traits among romantic couples. *Journal of Research in Personality*, *44*(3), 353–362. https://doi.org/10.1016/j.jrp.2010.03.008

Letzring, T. D., Wells, S. M. & Funder, D. C. (2006). Information quantity and quality affect the realistic accuracy of personality judgment. *Journal of Personality and Social Psychology*, *91*(1), 111–123. https://doi.org/10.1037/0022-3514.91.1.111

Levashina, J., Hartwell, C. J., Morgeson, F. P. & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*, *67*(1), 241–293. https://doi.org/10.1111/peps.12052

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

Lippa, R. A. & Dietz, J. K. (2000). The relation of gender, personality, and intelligence to judges' accuracy in judging strangers' personality from brief video segments. *Journal of Nonverbal Behavior*, *24*(1), 25–43. https://doi.org/10.1023/A:1006610805385

Little, T. D., Cunningham, W. A., Shahar, G. & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, *9*(2), 151–173. https://doi.org/10.1207/S15328007SEM0902_1

Loe, B. S., Sun, L., Simonfy, F. & Doebler, P. (2018). Evaluating an automated number series item generator using linear logistic test models. *Journal of Intelligence*, *6*(2), Article 20. https://doi.org/10.3390/jintelligence6020020

MacCann, C., Duckworth, A. L. & Roberts, R. D. (2009). Empirical identification of the major facets of conscientiousness. *Learning and Individual Differences*, *19*(4), 451–458. https://doi.org/10.1016/j.lindif.2009.03.007

MacCann, C., Joseph, D. L., Newman, D. A. & Roberts, R. D. (2014). Emotional intelligence is a second-stratum factor of intelligence: Evidence from hierarchical and bifactor models. *Emotion*, *14*(2), 358–374. https://doi.org/10.1037/a0034755

MacCann, C. & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion*, *8*(4), 540–551. https://doi.org/10.1037/a0012746

MacCann, C., Roberts, R. D., Matthews, G. & Zeidner, M. (2004). Consensus scoring and empirical option weighting of performance-based emotional intelligence (EI) tests. *Personality and Individual Differences*, *36*(3), 645–662. https://doi.org/10.1016/S0191-8869(03)00123-5

Markon, K. E., Quilty, L. C., Bagby, R. M. & Krueger, R. F. (2013). The development and psychometric properties of an informant-report form of the Personality Inventory for *DSM-5* (PID-5). *Assessment*, *20*(3), 370–383. https://doi.org/10.1177/1073191113486513

Matthews, G., Zeidner, M. & Roberts, R. D. (2012). Emotional intelligence: A promise unfulfilled? *Japanese Psychological Research*, *54*(2), 105–127. https://doi.org/10.1111/j.1468-5884.2011.00502.x

Mayer, J. D. (2008). Personal intelligence. *Imagination, Cognition and Personality*, *27*(3), 209–232. https://doi.org/10.2190/IC.27.3.b

Mayer, J. D. (2009). Personal intelligence expressed: A theoretical analysis. *Review of General Psychology*, *13*(1), 46–58. https://doi.org/10.1037/a0014229

Mayer, J. D. (2015). The personality systems framework: Current theory and development. *Journal of Research in Personality*, *56*, 4–14. https://doi.org/10.1016/j.jrp.2015.01.001

Mayer, J. D., Caruso, D. R. & Panter, A. T. (2019). Advancing the measurement of personal intelligence with the Test of Personal Intelligence, Version 5 (TOPI 5). *Journal of Intelligence*, *7*(1), Article 4. https://doi.org/10.3390/jintelligence7010004

Mayer, J. D., Caruso, D. R. & Salovey, P. (1999). Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, *27*(4), 267–298. https://doi.org/10.1016/S0160-2896(99)00016-1

Mayer, J. D., Caruso, D. R. & Salovey, P. (2016). The ability model of emotional intelligence: Principles and updates. *Emotion Review*, *8*(4), 290–300. https://doi.org/10.1177/1754073916639667

Mayer, J. D. & Geher, G. (1996). Emotional intelligence and the identification of emotion. *Intelligence*, *22*(2), 89–113. https://doi.org/10.1016/S0160-2896(96)90011-2

Mayer, J. D., Panter, A. T. & Caruso, D. R. (2012). Does personal intelligence exist? Evidence from a new ability-based measure. *Journal of Personality Assessment*, *94*(2), 124–140. https://doi.org/10.1080/00223891.2011.646108

Mayer, J. D., Panter, A. T. & Caruso, D. R. (2017). A closer look at the Test of Personal Intelligence (TOPI). *Personality and Individual Differences*, *111*, 301–311. https://doi.org/10.1016/j.paid.2017.02.008

Mayer, J. D. & Salovey, P. (1997). What is emotional intelligence? In P. Salovey & D. J. Sluyter (Eds.), *Emotional development and emotional intelligence: Educational implications* (pp. 3–34). Basic Books.

Mayer, J. D., Salovey, P., Caruso, D. R. & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion, 3*(1), 97–105. https://doi.org/10.1037/1528-3542.3.1.97

McCrae, R. R. (1994). The counterpoint of personality assessment: Self reports and observer ratings. *Assessment, 1*(2), 159–172. https://doi.org/10.1177/1073191194001002006

McCrae, R. R. & Costa, P. T., Jr. (1989). The structure of interpersonal traits: Wiggins's circumplex and the five-factor model. *Journal of Personality and Social Psychology*, *56*(4), 586–595. https://doi.org/10.1037/0022-3514.56.4.586

McCrae, R. R. & Costa P. T., Jr. (2003). *Personality in adulthood: A five-factor theory perspective* (2nd ed.). The Guilford Press.

McCrae, R. R. & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

McDaniel, M. A., Psotka, J., Legree, P. J., Yost, A. P. & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology*, *96*(2), 327–336. https://doi.org/10.1037/a0021983

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

Melby-Lervåg, M. & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology*, *49*(2), 270–291. https://doi.org/10.1037/a0028228

Millisecond Software. (2016). *Inquisit 4* [Computer software]. https://www.millisecond.com

Mischel, W. & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*(2), 246–268. https://doi.org/10.1037/0033-295x.102.2.246

Mischel, W. & Shoda, Y. (2008). Toward a unified theory of personality: Integrating dispositions and processing dynamics within the cognitive-affective processing system. In O. P. John, R. W. Robins & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 208–241). The Guilford Press.

Moutafi, J., Furnham, A. & Paltiel, L. (2004). Why is conscientiousness negatively correlated with intelligence? *Personality and Individual Differences*, *37*(5), 1013–1022. https://doi.org/10.1016/j.paid.2003.11.010

Murphy, N. A. (2012). Nonverbal perception. In S. T. Fiske & C. N. Macrae (Eds.), *The SAGE handbook of social cognition* (pp. 191–210). SAGE Publications. https://doi.org/10.4135/9781446247631.n10

Murphy, N. A. & Hall, J. A. (2011). Intelligence and interpersonal sensitivity: A meta-analysis. *Intelligence*, *39*(1), 54–63. https://doi.org/10.1016/j.intell.2010.10.001

Muthén, L. K. & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Muthén, L. K. & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

Naumann, L. P., Vazire, S., Rentfrow, P. J. & Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and Social Psychology Bulletin*, *35*(12), 1661–1671. https://doi.org/10.1177/0146167209346309

Nederström, M. & Salmela-Aro, K. (2014). Self-other agreement of personality judgments in job interviews: Exploring the effects of trait, gender, age and social desirability. *Scandinavian Journal of Psychology*, *55*(5), 520–526. https://doi.org/10.1111/sjop.12154

Nestler, S. & Back, M. D. (2013). Applications and extensions of the lens model to understand interpersonal judgments at zero acquaintance. *Current Directions in Psychological Science*, *22*(5), 374–379. https://doi.org/10.1177/0963721413486148

Nestler, S. & Back, M. D. (2017). Using cross-classified structural equation models to examine the accuracy of personality judgments. *Psychometrika*, *82*(2), 475–497. https://doi.org/10.1007/s11336-015-9485-6

Neyer, F. J. & Asendorpf, J. B. (2018). *Psychologie der Persönlichkeit*. Springer. https://doi.org/10.1007/978-3-662-54942-1

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, *66*(6), 574–583. https://doi.org/10.1037/h0040291

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments & Computers*, *32*(3), 396–402. https://doi.org/10.3758/BF03200807

Oltmanns, T. F. & Turkheimer, E. (2009). Person perception and personality pathology. *Current Directions in Psychological Science*, *18*(1), 32–36. https://doi.org/10.1111/j.1467-8721.2009.01601.x

Orchard, B., MacCann, C., Schulze, R., Matthews, G., Zeidner, M. & Roberts, R. D. (2009). New directions and alternative approaches to the measurement of emotional intelligence. In C. Stough, D. H. Saklofske & J. D. A. Parker (Eds..), *Assessing emotional intelligence: Theory, research, and applications* (pp. 321–344). Springer. https://doi.org/10.1007/978-0-387-88370-0_17

Ostendorf, F. (n.d.). *German translation of the 50-Item lexical Big-Five factor markers*. International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality and Other Individual Differences. https://ipip.ori.org/German50-itemBigFiveFactorMarkers.htm

Ostendorf, F. (1990). *Sprache und Persönlichkeitsstruktur: Zur Validität des Fünf-Faktoren-Modells der Persönlichkeit*. S. Roderer Verlag.

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

O'Sullivan, M. & Guilford, J. P. (1975). Six factors of behavioral cognition: Understanding other people. *Journal of Educational Measurement*, *12*(4), 255–271. https://doi.org/10.1111/j.1745-3984.1975.tb01027.x

O'Sullivan, M., Guilford, J. P. & deMille, R. (1965). *The measurement of social intelligence* (ED010278). ERIC. http://files.eric.ed.gov/fulltext/ED010278.pdf

Paunonen, S. V. & Hong, R. Y. (2013). The many faces of assumed similarity in perceptions of personality. *Journal of Research in Personality*, *47*(6), 800–815. https://doi.org/10.1016/j.jrp.2013.08.007

Piedmont, R. L. (1998). *The revised NEO Personality Inventory: Clinical and research applications.* Springer. https://doi.org/10.1007/978-1-4899-3588-5

Pisters, M. & Schulze, R. (2017, September 4–6). *Der Einfluss der Scoring-Methode auf die psychometrische Qualität von Testverfahren zur Erfassung von Emotionaler Intelligenz* [Forschungsreferat]. 14. Arbeitstagung der Fachgruppe Differentielle Psychologie, Persönlichkeitspsychologie und Psychologische Diagnostik, München.

Powell, D. M. & Bourdage, J. S. (2016). The detection of personality traits in employment interviews: Can "good judges" be trained? *Personality and Individual Differences*, *94*, 194–199. https://doi.org/10.1016/j.paid.2016.01.009

Powell, D. M. & Goffin, R. D. (2009). Assessing personality in the employment interview: The impact of training on rater accuracy. *Human Performance*, *22*(5), 450–465. https://doi.org/10.1080/08959280903248450

Preacher, K. J. & Coffman, D. L. (2006). *Computing power and minimum sample size for RMSEA* [Computer software]. http://www.quantpsy.org/

Primi, R. (2001). Complexity of geometric inductive reasoning tasks Contribution to the understanding of fluid intelligence. *Intelligence*, *30*(1), 41–70. https://doi.org/10.1016/S0160-2896(01)00067-8

Rauthmann, J. F., Gallardo-Pujol, D., Guillaume, E. M., Todd, E., Nave, C. S., Sherman, R. A., Ziegler, M., Jones, A. B. & Funder, D. C. (2014). The Situational Eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, *107*(4), 677–718. https://doi.org/10.1037/a0037250

Rauthmann, J. F. & Sherman, R. A. (2016). Measuring the Situational Eight DIAMONDS characteristics of situations: An optimization of the RSQ-8 to the S8*. *European Journal of Psychological Assessment*, *32*(2), 155–164. https://doi.org/10.1027/1015-5759/a000246

Ready, R. E., Clark, L. A., Watson, D. & Westerhouse, K. (2000). Self- and peer-reported personality: Agreement, trait ratability, and the "self-based heuristic". *Journal of Research in Personality*, *34*(2), 208–224. https://doi.org/10.1006/jrpe.1999.2280

Reallusion. (2014). *iClone 6* [Computer software]. https://www.reallusion.com

Riggio, R. E., Messamer, J. & Throckmorton, B. (1991). Social and academic intelligence: Conceptually distinct but overlapping constructs. *Personality and Individual Differences*, *12*(7), 695–702. https://doi.org/10.1016/0191-8869(91)90225-Z

Rikoon, S. H., Brenneman, M., Kim, L. E., Khorramdel, L., MacCann, C., Burrus, J. & Roberts, R. D. (2016). Facets of conscientiousness and their differential relationships with cognitive ability factors. *Journal of Research in Personality*, *61*, 22–34. https://doi.org/10.1016/j.jrp.2016.01.002

Roberts, R. D., Zeidner, M. & Matthews, G. (2001). Does emotional intelligence meet traditional standards for an intelligence? Some new data and conclusions. *Emotion*, *1*(3), 196–231. https://doi.org/10.1037/1528-3542.1.3.196

Rogers, K. H. & Biesanz, J. C. (2019). Reassessing the good judge of personality. *Journal of Personality and Social Psychology*, *117*(1), 186–200. https://doi.org/10.1037/pspp0000197

Romney, D. M. & Pyryt, M. C. (1999). Guilford's concept of social intelligence revisited. *High Ability Studies*, *10*(2), 137–142. https://doi.org/10.1080/1359813990100202

Roseman, I. J. (2001). A model of appraisal in the emotion system: Integrating theory, research, and applications. In K. R. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 68–91). Oxford University Press.

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L. & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test.* The Johns Hopkins University Press.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10, pp. 173–220). Academic Press. https://doi.org/10.1016/S0065-2601(08)60357-3

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Satorra, A. & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. https://doi.org/10.1007/BF02296192

Satorra, A. & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243–248. https://doi.org/10.1007/s11336-009-9135-y

Saucier, G. & Goldberg, L. R. (1996). The language of personality: Lexical perspectives on the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 21–50). The Guilford Press.

Saucier, G. & Ostendorf, F. (1999). Hierarchical subcomponents of the Big Five personality factors: A cross-language replication. *Journal of Personality and Social Psychology*, *76*(4), 613–627. https://doi.org/10.1037//0022-3514.76.4.613

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1–2), 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5

Schlegel, K., Boone, R. T. & Hall, J. A. (2017). Individual differences in interpersonal accuracy: A multi-level meta-analysis to assess whether judging other people is one skill or many. *Journal of Nonverbal Behavior*, *41*(2), 103–137. https://doi.org/10.1007/s10919-017-0249-0

Schmid Mast, M., Bangerter, A., Bulliard, C. & Aerni, G. (2011). How accurate are recruiters' first impressions of applicants in employment interviews? *International Journal of Selection and Assessment*, *19*(2), 198–208. https://doi.org/10.1111/j.1468-2389.2011.00547.x

Schneider, W. J. & McGrew, K. S. (2018). The Cattell–Horn–Carroll theory of cognitive abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 73–163). The Guilford Press.

Schreiber, M. & Iller, M.-L. (2016). *Handbuch Fragebogen zur Erfassung der Persönlichkeit (IPIP-240)*. IAP Institut für Angewandte Psychologie.

Schultze, M. (2018). *stuart: Subtests using algorithmic rummaging techniques* (R package version 0.7.3) [Computer software]. https://cran.r-project.org/package=stuart

Schulze, R., Holtzman, S., MacCann, C. & Roberts, R. D. (2009, September 16–19). *Assessment of emotional understanding with the Empathic Agent Paradigm (EAP)* [Paper presentation]. Tenth European Conference on Psychological Assessment, Ghent, Belgium.

Schulze, R. & Jobmann, A.-L. (2016, September 18–22). *Der Zusammenhang zwischen Emotional Understanding und Arbeitsgedächtniskapazität* [Forschungsreferat]. 50. Kongress der Deutschen Gesellschaft für Psychologie, Leipzig.

Schulze, R. & Roberts, R. D. (2006). Assessing the Big Five: Development and validation of the Openness Conscientiousness Extraversion Agreeableness Neuroticism Index Condensed (OCEANIC). *Zeitschrift Für Psychologie*, *214*(3), 133–149. https://doi.org/10.1026/0044-3409.214.3.133

Schulze, R. & Roberts, R. D. (2015). *The Acquisition-Application Test Design Principle (AcquA) for the generation of performance tests* [Manuscript in preparation]. Chair of Methods and Psychological Assessment, University of Wuppertal.

Seidel, K. (2007). *Social intelligence and auditory intelligence—Useful constructs?* [Doctoral dissertation, Otto-von-Guericke-University Magdeburg]. Share_it. http://dx.doi.org/10.25673/3851

Sherman, R. A., Nave, C. S. & Funder, D. C. (2010). Situational similarity and personality predict behavioral consistency. *Journal of Personality and Social Psychology*, *99*(2), 330–343. https://doi.org/10.1037/a0019796

Shoda, Y., Mischel, W. & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology*, *67*(4), 674–687. https://doi.org/10.1037/0022-3514.67.4.674

Snyder, M. & Ickes, W. (1985). Personality and social behavior. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (3rd ed., Vol. 2, pp. 883–947). Random House.

Soto, C. J. (2016). The little six personality dimensions from early childhood to early adulthood: Mean-level age and gender differences in parents' reports. *Journal of Personality*, *84*(4), 409–422. https://doi.org/10.1111/jopy.12168

Soto, C. J. & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Speer, A. B., Christiansen, N. D. & Laginess, A. J. (2019). Social intelligence and interview accuracy: Individual differences in the ability to construct interviews and rate accurately. *International Journal of Selection and Assessment*, *27*(2), 104–128. https://doi.org/10.1111/ijsa.12237

Srivastava, S. (2010). The five-factor model describes the structure of social perceptions. *Psychological Inquiry*, *21*(1), 69–75. https://doi.org/10.1080/10478401003648815

Stewart, T. R. (2001). The lens model equation. In K. R. Hammond & T. R. Stewart (Eds.), *The essential Brunswik: Beginnings, explications, applications* (pp. 357–362). Oxford University Press.

Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O. & Schulze, R. (2002). Working-memory capacity explains reasoning ability—and a little bit more. *Intelligence*, *30*(3), 261–288. https://doi.org/10.1016/S0160-2896(01)00100-3

Süß, H.-M., Seidel, K., Weis, S., Baumgarten, M., Karthaus, C., Nötzold, J. & Strien, J. (2009). *Magdeburger Test zur Sozialen Intelligenz (MTSI-3)* [Unveröffentlichtes Testverfahren]. Abteilung für Psychologische Methodenlehre, Psychodiagnostik und Evaluationsforschung, Institut für Psychologie, Otto-von-Guericke Universität Magdeburg.

Tackett, J. L. (2011). Parent informants for child personality: Agreement, discrepancies, and clinical utility. *Journal of Personality Assessment*, *93*(6), 539–544. https://doi.org/10.1080/00223891.2011.608763

Taft, R. (1955). The ability to judge people. *Psychological Bulletin*, *52*(1), 1–23. https://doi.org/10.1037/h0044999

Thielmann, I., Hilbig, B. E. & Zettler, I. (2020). Seeing me, seeing you: Testing competing accounts of assumed similarity in personality judgments. *Journal of Personality and Social Psychology*, *118*(1), 172–198. https://doi.org/10.1037/pspp0000222

Thielmann, I., Rau, R. & Locke, K. D. (2022). Trait-specificity versus global positivity: A critical test of alternative sources of assumed similarity in personality judgments. *Journal of Personality and Social Psychology*. Advance online publication. https://doi.org/10.1037/pspp0000420

Thurstone, L. L. (1938). *Primary mental abilities*. University of Chicago Press.

Tong, S. T., Corriero, E. F., Wibowo, K. A., Makki, T. W. & Slatcher, R. B. (2020). Self-presentation and impressions of personality through text-based online dating profiles: A lens model analysis. *New Media & Society*, *22*(5), 875–895. https://doi.org/10.1177/1461444819872678

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review*, *93*(3), 239–257. https://doi.org/10.1037/0033-295X.93.3.239

Trope, Y. & Liberman, A. (1993). The use of trait conceptions to identify other people's behavior and to draw inferences about their personalities. *Personality and Social Psychology Bulletin*, *19*(5), 553–562. https://doi.org/10.1177/0146167293195007

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review*, *71*(6), 528–530. https://doi.org/10.1037/h0047061

Tupes, E. C. & Christal, R. E. (1992). Recurrent personality factors based on trait ratings. *Journal of Personality*, *60*(2), 225–251. https://doi.org/10.1111/j.1467-6494.1992.tb00973.x

van der Linden, D., te Nijenhuis, J. & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of big five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, *44*(3), 315–327. https://doi.org/10.1016/j.jrp.2010.03.003

Vernon, P. E. (1933). Some characteristics of the good judge of personality. *The Journal of Social Psychology*, *4*(1), 42–57. https://doi.org/10.1080/00224545.1933.9921556

Vogt, D. S. & Colvin, C. R. (2003). Interpersonal orientation and the accuracy of personality judgements. *Journal of Personality*, *71*(2), 267–295. https://doi.org/10.1111/1467-6494.7102005

Waterhouse, L. (2006a). Multiple intelligences, the Mozart effect, and emotional intelligence: A critical review. *Educational Psychologist*, *41*(4), 207–225. https://doi.org/10.1207/s15326985ep4104_1

Waterhouse, L. (2006b). Inadequate evidence for multiple intelligences, Mozart effect, and emotional intelligence theories. *Educational Psychologist*, *41*(4), 247–255. https://doi.org/10.1207/s15326985ep4104_5

Watrin, L., Hülür, G. & Wilhelm, O. (2022). Training working memory for two years—no evidence of transfer to intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(5), 717–733. https://doi.org/10.1037/xlm0001135

Watson, D. & Clark, L. A. (1997). Extraversion and its positive emotional core. In R. Hogan, J. A. Johnson & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 767–793). Academic Press. https://doi.org/10.1016/B978-012134645-4/50030-5

Watson, D., Hubbard, B. & Wiese, D. (2000). Self–other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of Personality and Social Psychology*, *78*(3), 546–558. https://doi.org/10.1037/0022-3514.78.3.546

Wedeck, J. (1947). The relationship between personality and „psychological ability“. *British Journal of Psychology*, *37*(3), 133–151. https://doi.org/10.1111/j.2044-8295.1947.tb01128.x

Weekley, J. A., Ployhart, R. E. & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 157–182). Lawrence Erlbaum Associates Publishers.

Weis, S. (2008). *Theory and measurement of social intelligence as a cognitive performance construct* [Doctoral dissertation, Otto-von-Guericke-University Magdeburg]. Share_it. http://dx.doi.org/10.25673/4842

Weis, S., Seidel, K. & Süß, H.-M. (2006). Messkonzepte sozialer Intelligenz – Literaturübersicht und Ausblick. In R. Schulze, P. A. Freund & R. D. Roberts (Hrsg.), *Emotionale Intelligenz: Ein internationales Handbuch* (S. 213–234). Hogrefe.

Weis, S. & Süß, H.-M. (2005). Social intelligence—A review and critical discussion of measurement concepts. In R. Schulze & R. D. Roberts (Eds.), *Emotional intelligence: An international handbook* (pp. 203–230). Hogrefe & Huber Publishers.

Weis, S. & Süß, H.-M. (2007). Reviving the search for social intelligence – A multitrait-multimethod study of its structure and construct validity. *Personality and Individual Differences*, *42*(1), 3–14. https://doi.org/10.1016/j.paid.2006.04.027

West, T. V. & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review*, *118*(2), 357–378. https://doi.org/10.1037/a0022936

Woods, S. A. & Anderson, N. R. (2016). Toward a periodic table of personality: Mapping personality scales between the five-factor model and the circumplex model. *Journal of Applied Psychology*, *101*(4), 582–604. https://doi.org/10.1037/apl0000062

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]. https://www.statmodel.com/download/Yudissertation.pdf

Zajenkowski, M. & Stolarski, M. (2015). Is conscientiousness positively or negatively related to intelligence? Insights from the national level. *Learning and Individual Differences*, *43*, 199–203. https://doi.org/10.1016/j.lindif.2015.08.009

Zebrowitz, L. A. & Collins, M. A. (1997). Accurate social perception at zero acquaintance: The affordances of a Gibsonian approach. *Personality and Social Psychology Review*, *1*(3), 204–223. https://doi.org/10.1207/s15327957pspr0103_2