Selbsteinschätzungen sprachlicher Kompetenzen im Deutschen jugendlicher Flüchtlinge

INAUGURAL-DISSERTATION
in der Fakultät Humanwissenschaften
der Otto-Friedrich-Universität Bamberg

vorgelegt von
Anike Schild, geb. Dröscher
aus Kirn

Bamberg 2024

BAMBERG
GRADUATE SCHOOL
OF SOCIAL SCIENCES

Tag der mündlichen Prüfung: 01.12.2023

Dekan: Universitätsprofessor Dr. Claus-Christian Carbon

Betreuerin: Universitätsprofessorin Dr. Cordula Artelt

Weitere Gutachterin: Universitätsprofessorin Dr. Sabine Weinert

URN: urn:nbn:de:bvb:473-irb-943495
DOI: https://doi.org/10.20378/irb-94349

Danksagung

Zum Gelingen dieser Arbeit haben viele liebe Menschen beigetragen, denen ich an dieser Stelle gerne danken möchte.

Meiner Erstbetreuerin Prof. Dr. Cordula Artelt danke ich dafür, dass sie sich regelmäßig die Zeit genommen hat, mir mit sehr wertvollem Feedback und Denkanstößen zu helfen, die Qualität meiner Arbeit zu verbessern und mich wissenschaftlich weiterzuentwickeln.

Prof. Dr. Sabine Weinert danke ich dafür, dass sie sich bereit erklärt hat, die Zweitbegutachtung zu übernehmen und dass sie im Rahmen des Forschungsseminars konstruktive Diskussionen zu meiner Arbeit angestoßen hat.

Dr. Gisela Will, Koordinatorin des ReGES-Projekts am LIfBi, danke ich für die Unterstützung meiner Promotion und den Gestaltungsspielraum, den sie mir bei meiner Arbeit im Projekt gegeben hat, sowie für den konstruktiven Austausch bei der Entwicklung der Fragebogeninhalte, die ich für meine Dissertation nutzen konnte.

Dem gesamten ReGES-Projektteam danke ich für den Zusammenhalt und die gegenseitige Unterstützung bei der Projektarbeit sowie für den anregungsreichen Austausch zur Promotion. Dabei danke ich insbesondere Ebru Balaban-Feldens und Jörg Welker dafür, dass sie unsere langjährige gemeinsame Zeit im ReGES-Projekt und den Weg zur Promotion sehr bereichert haben und dass sie in unseren motivierenden Promotions-Austauschtreffen immer ein offenes Ohr und hilfreiche Tipps hatten.

Dr. Timo Gnambs, Dr. Kathrin Lockl und Dr. Lena Nusser danke ich für den Austausch zu statistischen Fragen in der Forschungssprechstunde.

Meiner Freundin Dr. Theresa Fehn danke ich für das aufmerksame und kritische Korrekturlesen und die konstruktiven Verbesserungsvorschläge zu meiner Doktorarbeit.

Meiner Freundin Dr. Katrin Wolstein danke ich für das aufmerksame und bestärkende Korrekturlesen und die konstruktiven Verbesserungsvorschläge und dafür, dass sie mit mir durch alle Höhen und Tiefen der Promotionsphase geht.

Meinen Schwiegereltern Beate und Bernhard Schild danke ich für die liebevolle regelmäßige Betreuung meines Sohnes, ohne die die Doktorarbeit noch nicht fertig wäre.

Ebenso danke ich meinen Eltern Sabine Heil und Julian Dröscher, die trotz der räumlichen Entfernung immer wieder bei der Kinderbetreuung unterstützt und mitangepackt haben. Aber insbesondere danke ich meinen Eltern und meiner Familie für alles, was sie mir mit auf den Weg gegeben haben, dass sie mir immer Rückhalt geben und alles unterstützen, was ich tue.

Meinem Mann Sebastian Schild danke ich für alles, was wir in den vergangenen Jahren gemeinsam geschafft haben und dass ich mich immer auf ihn verlassen kann.

Meinen Kindern Emilian und Elea danke ich für ihre Geduld und dafür, dass sie die besten Motivatoren sind, die Promotion zielstrebig abzuschließen.

Für die finanzielle Unterstützung durch ein Promotionsabschlussstipendium bedanke ich mich bei der Bamberg Graduate School of Social Sciences.

Vorwort

Bereits in der Masterarbeit habe ich mich mit der Diagnostik kognitiver Fähigkeiten bei Geflüchteten beschäftigt. Als im Sommer 2016 die vom BMBF geförderte Flüchtlingsstudie Refugees in the German Educational System (ReGES) am Leibniz-Institut für Bildungsverläufe in Bamberg startete, bin ich dort als wissenschaftliche Mitarbeiterin eingestiegen. Schwerpunkte meiner Arbeit waren die Kompetenztestungen zu kognitiven Grundfähigkeiten und Sprachkompetenzen sowie das Fragebogenmodul zum Thema Sprache. So hatte ich die Gelegenheit, verschiedene Arten von Selbsteinschätzungsitems zu den Sprachkompetenzen der jugendlichen Geflüchteten in die Befragungen einzubringen, geeignete Tests deutscher Sprachfähigkeiten mit auszuwählen und den gesamten Prozess von der Vorbereitung der Erhebungen bis hin zur Aufbereitung der Kompetenzdaten mitumzusetzen. Die Zusammenstellung der Selbsteinschätzungsitems weckte mein Interesse daran, wie genau wir die Sprachkompetenzen der Jugendlichen auf diese Weise erfassen könnten. Durch die Gestaltungsmöglichkeit und die Kenntnis des Projekts und der Daten hatte ich die optimale Grundlage, diese Daten für meine Doktorarbeit zu nutzen und mir das Ergebnis der Selbsteinschätzungen der jugendlichen Flüchtlinge genau anzuschauen. Ich freue mich, meine gewonnenen Erkenntnisse hier zu teilen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Gegenstand dieser Arbeit sind die Selbsteinschätzungen sprachlicher Kompetenzen im Deutschen jugendlicher Flüchtlinge. Im Fokus steht, wie genau die jugendlichen Flüchtlinge ihre sprachlichen Kompetenzen im Deutschen einschätzen und ob sich die Selbsteinschätzungen als Maß für die sprachlichen Kompetenzen im Deutschen eignen. Es werden Faktoren identifiziert, die Selbsteinschätzungen zusätzlich zu der tatsächlichen sprachlichen Kompetenz systematisch beeinflussen und es werden verschiedene Arten von Selbsteinschätzungsitems hinsichtlich ihrer Eignung zur Erfassung der sprachlichen Kompetenzen im Deutschen verglichen. Betrachtet werden vier verschiedene Arten von Selbsteinschätzungsitems zum Verstehen und Sprechen der deutschen Sprache. Dazu gehören 1) allgemein formulierte Standard-Items, die eine fünfstufige Antwortskala vorgeben, 2) allgemein formulierte Schieberegler-Items, die einen zehnstufigen Endpunkt-gelabelten Schieberegler als Antwortskala vorgeben, 3) allgemein formulierte Vergleich-Items, die gleichaltrige Muttersprachlerinnen und Muttersprachler als Referenzgruppe vorgeben mit einer fünfstufigen Antwortskala und 4) spezifisch formulierte Can-Do-Statements, von denen diejenigen Anforderungen ausgewählt werden sollen, die die Teilnehmenden auf Deutsch können. Als Kriterium werden die Ergebnisse zweier Deutschkompetenztests, des *Peabody-Picture-Vocabulary-Tests – 4. Ausgabe* (PPVT-4; Lenhard et al., 2015) zur Erfassung des rezeptiven Wortschatzes und des *Tests zur Überprüfung des Grammatikverständnisses* (TROG-D; Fox-Boyer, 2016), herangezogen. Analysiert werden Daten von 1 877 Teilnehmenden der ersten Erhebungswelle und von 778 Teilnehmenden der siebten Erhebungswelle der Studie *Refugees in the German Educational System* (ReGES). Die Teilnehmenden sind als Geflüchtete nach Deutschland gekommen und waren im Durchschnitt in der ersten Erhebungswelle 16.0 Jahre und in der siebten Erhebungswelle 17.8 Jahre alt. Die Korrelationen zwischen den Selbsteinschätzungen und den Deutschkompetenztests weisen mittlere Effektstärken auf und die jugendlichen Flüchtlinge überschätzten ihre Deutschkompetenzen im Durchschnitt. Als Antwort auf die allgemein formulierten Items wählten die allermeisten Jugendlichen nur die wenigen positiv formulierten Kategorien, sodass die Verteilung der Antworten schief war, teilweise mit deutlichem Deckeneffekt und die breite Variation der einzuschätzenden Kompetenz nicht differenziert abgebildet wurde. Mithilfe von Strukturgleichungsmodellen werden Einflussfaktoren der Selbsteinschätzungen identifiziert und dabei der Einfluss der objektiv gemessenen Kompetenz auf die Selbsteinschätzungen kontrolliert. Die Fähigkeit zum schlussfolgernden Denken hatte einen negativen Effekt auf die Selbsteinschätzungen, sodass Teilnehmende mit höheren Fähigkeiten zum schlussfolgernden Denken ihre Deutschkompetenzen vergleichsweise niedriger einschätzten als Teilnehmende mit niedrigeren Fähigkeiten zum schlussfolgernden Denken. Das Engagement beim Deutschlernen hatte einen positiven Effekt auf die Selbsteinschätzungen, sodass Personen, die mehr Engagement beim Deutschlernen angaben, ihre

Deutschkompetenzen vergleichsweise höher einschätzten als Personen, die weniger Engagement beim Deutschlernen angaben. Der vorhergesagte negative Effekt der Leistung in Mathematik wurde nicht bestätigt, ebenso wenig haben sich die vorhergesagten negativen Effekte der Teilnahme an einem Deutschkurs und der Teilnahme an einem Deutschtest auf die Höhe der Selbsteinschätzung bestätigt. Die verschiedenen Arten von Selbsteinschätzungsitems unterschieden sich nicht grundlegend in der Genauigkeit, mit der sie die deutschen Sprachkompetenzen erfassten. Anhand eines Latent-Change-Modells wird gezeigt, dass die Veränderung der sprachlichen Kompetenzen im Deutschen zwischen den beiden Messzeitpunkten mit den Can-Do-Statements besser erfasst werden kann als mit den Standard-Items. Es wird geschlussfolgert, dass die Selbsteinschätzungen die Deutschkompetenzen der Teilnehmenden nur ungenau erfassen und objektive Kompetenzmaße grundsätzlich zu bevorzugen sind. Sofern Selbsteinschätzungen als Deutschkompetenzmaße herangezogen werden, sollten bei der Interpretation der Ergebnisse die Ungenauigkeiten und systematischen Einflüsse z.B. der Fähigkeit zum schlussfolgernden Denken oder des Engagements beim Deutschlernen berücksichtigt werden. Möglichkeiten zur Verbesserung der Selbsteinschätzungsitems mit dem Ziel einer genaueren Erfassung der Deutschkompetenzen werden diskutiert.

Abstract

Subject of this dissertation are the self-assessments of German language competences of adolescent refugees. The focus is on how accurately the adolescent refugees assess their German language competences and whether the self-assessments are suitable as a measure of German language competences. Factors that systematically influence the self-assessments in addition to the actual language competence are identified, and different kinds of self-assessment items are compared regarding their suitability to measure German language competences. Four different kinds of self-assessment items on understanding and speaking the German language are considered. These include 1) generally formulated standard items with a five-point response scale, 2) generally formulated slider items with a ten-point endpoint-labeled slider as the response scale, 3) generally formulated comparison items that specify native speakers of the same age as a reference group with a five-point response scale, and 4) specifically formulated can-do statements, from which those requirements that the participants can do in German are to be selected. As a criterion, results of two German language tests are used: the German version of the *Peabody-Picture-Vocabulary-Test – 4th edition* (PPVT-4; Lenhard et al., 2015) to assess receptive vocabulary, and the German version of the *Test for Reception of Grammar* (TROG-D; Fox-Boyer, 2016). Data of 1 877 participants of the first wave and 778 participants of the seventh wave of the study *Refugees in the German Educational System* (ReGES) are analyzed. Participants came to Germany as refugees and were on average 16.0 years old in the first wave and 17.8 years old in the seventh wave. Correlations between self-assessments and test scores of German language competence show medium effect sizes and the adolescent refugees overestimated their German proficiency on average. Most adolescents chose only the few positively formulated categories in response to the generally formulated items, which resulted in skewed distributions of the responses, strong ceiling effects for some kinds of items, and the broad variation of the competence to be assessed was not represented in a differentiated manner. Structural equation models are used to identify factors influencing the self-assessments, controlling for the influence of the objectively measured competence on the self-assessments. Reasoning ability had a negative effect on the self-assessments, i.e., participants with higher reasoning abilities estimated their German language competence to be comparatively lower than participants with lower reasoning abilities. Involvement in learning German had a positive effect on the self-assessments, i.e., participants who reported more involvement in learning German rated their German skills comparatively higher than people who reported less involvement in learning German. The predicted negative effect of mathematical achievement on the self-assessments was not confirmed, nor were the predicted negative effects of participation in a German course and participation in a German language test on the level of self-assessment confirmed. The different kinds of self-assessment items did not differ substantially considering how accurately they measured German language

competences. Analyzing a latent change score model, it is shown that the change of the German language competence between the two measurement time points was better represented by the can-do statements than by the standard items. I conclude that self-assessments measure German language competences only inaccurately and that objective measures are to be preferred in general. If self-assessments are however used as a measure of German language competence, inaccuracies and systematic influences of e.g. reasoning ability or the involvement in learning German should be considered when interpreting the results. Possibilities to improve self-assessment items to measure German language competences more accurately are discussed.

Einleitung

Kompetenzen in der Sprache des Aufnahmelandes sind ein Schlüssel zur Integration und wichtig für den Bildungserfolg junger Flüchtlinge (Edele et al., 2021; Esser, 2006). Insbesondere vor dem Hintergrund der stark gestiegenen Zuwanderung von Asylsuchenden in der Mitte des letzten Jahrzehnts nach Deutschland (Bundesamt für Migration und Flüchtlinge, 2020) und den Herausforderungen, die sich daraus für die Integration der jungen Flüchtlinge ins deutsche Bildungssystem ergeben haben, ist es eine wichtige Forschungsaufgabe, Erkenntnisse über die Deutschkompetenzen und die Entwicklung der Deutschkompetenzen junger Flüchtlinge zu gewinnen, um ggf. gezielte Maßnahmen zur Förderung der Deutschkompetenzen ergreifen zu können.

Die Erfassung von Sprachkompetenzen stellt in Studien jedoch eine Herausforderung dar: Die Implementation von geeigneten Kompetenztests ist aufwändig und teilweise kostspielig. Weiterhin dauert die Durchführung eines Kompetenztests vergleichsweise lange und kann somit in Konkurrenz zu anderen zu erhebenden Konstrukten stehen oder die Erhebungszeit verlängern, was eine mögliche Belastung für Teilnehmende darstellt. Weniger zeitintensiv, einfacher in der Durchführung und insgesamt ökonomischer ist es, Teilnehmende ihre Kompetenzen selbst einschätzen zu lassen (Herreen & Zajac, 2018; Neubauer & Hofer, 2019). Insbesondere für große Panelstudien ist die subjektive Kompetenzerfassung häufig die einzige Möglichkeit (Esser, 2006). Auch Studien zu Determinanten und Outcomes des Spracherwerbs von Migrantinnen und Migranten beruhen in der Regel auf Selbsteinschätzungen (Esser, 2006). Es hat sich jedoch gezeigt, dass der Zusammenhang zwischen Selbsteinschätzungen und objektiv gemessenen Kompetenzen stark variiert (Brantmeier et al., 2012; Freund & Kasten, 2012; Mabe & West, 1982; Ross, 1998; Zell & Krizan, 2014) und dass sich die Ergebnisse von Analysen zu Determinanten und Outcomes von Sprachkompetenzen unterscheiden, abhängig davon, ob in derselben Stichprobe die Selbsteinschätzung oder der Testscore als Sprachkompetenzmaß verwendet wird (Edele et al., 2015). Die Validität von Selbsteinschätzungen als Maß für Kompetenzen scheint also oft nicht zufriedenstellend zu sein und es ist umstritten, ob sie als Proxy bzw. als indirekte Messung der tatsächlichen Kompetenz verwendet werden können (Edele et al., 2015; Esser, 2006; Neubauer & Hofer, 2019). Selbsteinschätzungen haben demzufolge Vorteile in der Erhebung, aber es ist unklar, ob sie sich als Alternative für Kompetenztestungen eignen.

Um die Auswirkungen der Messung sprachlicher Kompetenzen im Deutschen mit subjektiven Selbsteinschätzungen anstelle von objektiven Kompetenztests auf Studienergebnisse abschätzen und berücksichtigen zu können, ist es zum einen wichtig zu wissen, wie valide die sprachlichen Kompetenzen mit Selbsteinschätzungsmaßen erfasst werden können. Die bisherigen Ergebnisse zur Validität von Selbsteinschätzungen sind jedoch heterogen und können sich abhängig von der

untersuchten Gruppe unterscheiden. Zudem wurden hauptsächlich korrelative Zusammenhänge zwischen Selbsteinschätzungen und anderen Kompetenzmaßen berechnet, nicht jedoch analysiert, ob die Sprachkompetenzen im Durchschnitt über- oder unterschätzt werden. Deshalb bleibt unklar, wie genau mit Selbsteinschätzungen die deutschen Sprachkompetenzen jugendlicher Flüchtlinge erfasst werden können. Folgende Eigenschaften der Gruppe der jugendlichen Flüchtlinge könnten sich darauf auswirken, wie genau sich deren sprachliche Kompetenzen mit Selbsteinschätzungen erfassen lassen: Die jugendlichen Flüchtlinge sind noch jung, was relevant für die Genauigkeit der Selbsteinschätzungen ist, da sich der altersabhängige kognitive Entwicklungsstand darauf auswirkt, wie Personen ihre Kompetenzen selbst einschätzen (vgl. Harter, 2012). Darüber hinaus könnten die unterschiedlichen kulturellen Hintergründe der jugendlichen Flüchtlinge die Art und Weise beeinflussen, wie diese sich selbst einschätzen, ob sie beispielsweise zu bescheidenen oder überhöhten Selbsteinschätzungen neigen (z.B. Chen et al., 1995; Min et al., 2016). Zuletzt liegt mit dem späten Zweitspracherwerb eine besondere Form des Spracherwerbs vor und die Erwerbsbedingungen sind individuell unterschiedlich. Die Jugendlichen haben in der Regel zunächst mindestens eine Erstsprache erworben und erst mit der Ankunft in Deutschland im späten Kindes- oder Jugendalter mit dem Deutscherwerb begonnen. Der Deutscherwerb erfolgte sowohl unsystematisch im Alltag, v.a. in der Schule, als auch systematisch in Sprachkursen in oder außerhalb der Schule. Deshalb gibt es keinen einheitlichen Referenzrahmen, an dem die jugendlichen Flüchtlinge ihre sprachlichen Kompetenzen messen können, wie es z.B. der Fall ist, wenn man Teilnehmende eines Fremdsprachkurses, in dem es einheitliche Lernziele gibt und in dem die anderen Teilnehmenden eine klar definierte Referenzgruppe darstellen, um eine Selbsteinschätzung bittet. Die jugendlichen Flüchtlinge unterscheiden sich demzufolge in für die Genauigkeit von Selbsteinschätzungen relevanten Merkmalen von anderen Stichproben, deren Selbsteinschätzungen von Kompetenzen empirisch untersucht wurden und die Genauigkeit ihrer Selbsteinschätzungen der Deutschkompetenz muss untersucht werden, um deren Eignung als Maße der Deutschkompetenzen zu beurteilen.

Um abschätzen zu können, welche Auswirkungen es hat, Selbsteinschätzungen als Maß für die Deutschkompetenzen jugendlicher Flüchtlinge zu verwenden, ist es darüber hinaus wichtig zu wissen, inwiefern die Selbsteinschätzungen systematischen Verzerrungen unterliegen. Verwendet man Selbsteinschätzungen von Sprachkompetenzen z.B., um Determinanten der Sprachkompetenz zu bestimmen, könnten Effekte über- oder unterschätzt werden, wenn die untersuchten Determinanten nicht nur die Sprachkompetenz beeinflussen, sondern darüber hinaus die Selbsteinschätzungen in eine bestimmte Richtung verzerren. Wenn z.B. Personen mit besseren Fähigkeiten zum schlussfolgernden Denken sowohl dazu neigen, ihre Deutschkompetenzen bei gleicher Deutschkompetenz niedriger einzuschätzen als Personen mit schlechteren Fähigkeiten zum schlussfolgernden Denken, aber die Fähigkeit zum schlussfolgernden Denken darüber hinaus mit

der tatsächlichen Deutschkompetenz so zusammenhängt, dass Personen mit besseren Fähigkeiten zum schlussfolgernden Denken die deutsche Sprache besser beherrschen, würde man den positiven Zusammenhang zwischen der Fähigkeit zum schlussfolgernden Denken und der deutschen Sprachkompetenz unterschätzen, wenn man die deutsche Sprachkompetenz nur mit Selbsteinschätzungen misst. Sind die Einflussfaktoren von Selbsteinschätzungen bekannt, können diese bei der Interpretation von Forschungsergebnissen, die auf Selbsteinschätzungen basieren, berücksichtigt werden.

Weiterhin könnten unterschiedliche Erhebungsformen von Selbsteinschätzungen unterschiedlich genau sein. Selbsteinschätzungen können auf vielfältige Art erhoben werden, mit unterschiedlichsten Items und Antwortskalen. Es gibt z.B. Hinweise, dass die Validität von Selbsteinschätzungen höher ist, wenn relative Skalen mit genau definierter Referenzgruppe vorgegeben werden (Freund & Kasten, 2012). Wenn bekannt ist, für welche Items unter gegebenen Bedingungen genauere Selbsteinschätzungen zu erwarten sind, können diese bevorzugt verwendet werden. Deshalb sollten auch verschiedene Selbsteinschätzungsmaße miteinander verglichen werden.

Neben der Messung der sprachlichen Kompetenzen im Deutschen zu einem bestimmten Zeitpunkt, könnte auch die Betrachtung der Veränderung der Deutschkompetenzen zwischen zwei oder mehr Zeitpunkten ein Forschungsziel sein. Auch wenn Selbsteinschätzungen in einer bestimmten Stichprobe zu einem gegebenen Messzeitpunkt eine valide Differenzierung hinsichtlich der Sprachkompetenz zulassen sollten, ist es möglich, dass sich die Entwicklung der Sprachkompetenz über die Zeit nicht mit Selbsteinschätzungen abbilden lässt. Verbessert sich die Sprachkompetenz mit der Zeit, müsste auch die Selbsteinschätzung ansteigen. Beeinflusst aber beispielsweise eine Referenzgruppe, deren Sprachkompetenz sich ebenfalls mit der Zeit verbessert, die Selbsteinschätzungen, würden diese möglicherweise nicht erwartungsgemäß steigen, sondern abhängig von dem Ergebnis des Vergleichs mit der Referenzgruppe auf dem gleichen Niveau wie zuvor bleiben. Dann würde die Verbesserung der Sprachkompetenz über die Zeit unterschätzt. Um beurteilen zu können, ob sich die Veränderung von Sprachkompetenzen über die Zeit mit Selbsteinschätzungen messen lässt, müssen grundlegende Kenntnisse über die Entwicklung von Selbsteinschätzungen von Sprachkompetenzen gewonnen werden.

Das Ziel dieser Dissertation ist es, Selbsteinschätzungen als Maß für die sprachlichen Kompetenzen im Deutschen jugendlicher Flüchtlinge umfassend zu untersuchen. Neben der Genauigkeit der Selbsteinschätzungen dieser Gruppe wird untersucht, welche individuellen Faktoren die Selbsteinschätzungen systematisch beeinflussen und wie genau verschiedenartige Items die sprachlichen Kompetenzen im Deutschen und deren Veränderung über die Zeit erfassen. Folglich wird der Forschungsstand zu Selbsteinschätzungen von sprachlichen Kompetenzen erweitert, insbesondere für die Gruppe der jugendlichen Flüchtlinge in Deutschland. Die Ergebnisse tragen zur

Interpretation von unter Verwendung von Selbsteinschätzungen gewonnenen Forschungsergebnissen sowie zu informationsbasierten Entscheidungen bei der Wahl eines Kompetenzmaßes bei.

1 Theoretischer und empirischer Hintergrund

Die Forschungsarbeiten, die zum Wissen über Selbsteinschätzungen beitragen, sind vielfältig, indem sie aus verschiedenen Fachrichtungen stammen, verschiedene Ansatzpunkte verfolgen und verschiedene Methoden anwenden. Sie bilden eine umfangreiche und wertvolle Basis, um ein tieferes Verständnis von Selbsteinschätzungen zu erlangen sowie Hypothesen zu deren weiteren Erforschung zu generieren. In diesem Kapitel wird ein Überblick über den für die Fragestellung der Dissertation relevanten theoretischen und empirischen Hintergrund gegeben.

1.1 Selbsteinschätzungen: Die Relevanz des Selbstkonzepts

Selbsteinschätzungen lassen sich im Kontext unseres informationsverarbeitenden Systems wie folgt einordnen. Zur Einschätzung einer Fähigkeit muss Wissen über die eigene Person im Zusammenhang mit dieser Fähigkeit aus einem kognitiven System abgerufen werden. Auf die eigene Person bezogenes deklaratives Wissen ist im zum Selbst gehörenden Selbstkonzept organisiert (z.B. Schütz et al., 2016). Eigene Fähigkeiten werden anhand der abgerufenen Informationen aus dem relevanten Bereich des Selbstkonzepts eingeschätzt und spiegeln Aspekte des Selbstkonzepts in einem bestimmten Bereich wider (vgl. Edele et al., 2015).

Als Grundlage für die weiteren Ausführungen zum Selbstkonzept wird dieses zunächst in das System des Selbst eingeordnet. „Das Selbst ist ein dynamisches System (Markus & Wurf, 1987), das einerseits auf die jeweilige Person bezogene Überzeugungs- und Erinnerungsinhalte in hochstrukturierter Form und andererseits die mit diesen Inhalten und Strukturen operierenden Prozesse und Mechanismen umfasst“ (Greve, 2000, S. 17). Grundlegend dafür, ein Selbst zu haben, ist die menschliche Fähigkeit zur Selbstreflektion (Leary & Tangney, 2012). Ende des 19. Jahrhunderts prägte William James (1890) die Unterscheidung zwischen *I-Selbst* und *Me-Selbst*. Das I-Selbst stellt die regulierende Instanz des Selbst dar, welche wahrnehmungs- und handlungssteuernde Funktionen übernimmt und als Subjekt der Erkenntnis fungiert („Self as knower“, übersetzt etwa „das erkennende Selbst“; Schütz et al., 2016, S. 145), während das Me-Selbst Objekt der Erkenntnis ist („Self as known“, übersetzt etwa „das erkannte Selbst“; Schütz et al., 2016, S. 145) und sich auf die „Wissensinhalte über sich selbst“ (Schütz et al., 2016, S. 145) bezieht. Neben affektiven (Selbstwert) und handlungsregulierenden (z.B. Selbstwirksamkeitserwartungen) Aspekten stellt das Selbstkonzept den kognitiven Aspekt des Me-Selbst dar (vgl. Schütz et al., 2016; Schütz et al., 2018).

Shavelson et al. (1976) definieren das Selbstkonzept grob als die Wahrnehmung einer Person von sich selbst und beschreiben es u.a. als facettenreich und hierarchisch. Das übergeordnete Selbstkonzept wird in das akademische und das nichtakademische Selbstkonzept unterteilt und

auf der nächsten Stufe des akademischen Selbstkonzepts wird zwischen dem verbalen und dem mathematischen Selbstkonzept unterschieden (Marsh & Shavelson, 1985). „Vorstellungen über die Höhe eigener Fähigkeiten“ werden auch als Fähigkeitsselbstkonzepte bezeichnet (Dickhäuser, 2006, S. 5). Nach Markus (1977) beinhaltet das Selbstkonzept verschiedene sogenannte Selbstschemata. In diesen werden selbstbezogene Informationen zu einzelnen Bereichen aus verschiedensten Quellen organisiert und gebündelt. Ein Selbstschema beeinflusst sowohl die Verarbeitung neuer selbstbezogener Informationen als auch das Verhalten in dem entsprechenden Bereich. Das Selbstkonzept dient dabei der Aufrechterhaltung eines kohärenten Selbstbilds (Harter, 2012; Swann & Buhrmester, 2012).

Nicht alle Informationen, die im Selbstkonzept organisiert sind, sind jederzeit gleichermaßen verfügbar. Das *Working-Self-Concept* stellt den aktiven Teil des Selbstkonzepts dar, der zu einem bestimmten Zeitpunkt verfügbar ist und die Wahrnehmung und das Verhalten beeinflusst. Welche Aspekte des Selbstkonzepts im Working-Self-Concept verfügbar sind, kann durch die Salienz dieser Aspekte manipuliert sein (Markus & Wurf, 1987). Das bedeutet, auch Selbsteinschätzungen spiegeln nicht unbedingt das gesamte gespeicherte Wissen über eine eigene Fähigkeit wider, sondern die Teile des Wissens, die zum Zeitpunkt der Einschätzung im Working-Self-Concept verfügbar sind.

Nachdem in diesem Abschnitt die Relevanz des Selbstkonzepts für die Selbsteinschätzungen herausgearbeitet wurde, wird im folgenden Kapitel 1.2 thematisiert, aus welchen Quellen die im Selbstkonzept gespeicherten Informationen gewonnen werden und welche Referenzen herangezogen werden, um diese zu bewerten.

1.2 Quellen und Referenzen für selbstbezogene Informationen

Selbstbezogenes Wissen kann aus verschiedenen Quellen gewonnen werden. Soziale Rückmeldungen und Reflektionen über mögliche Beurteilungen durch andere Personen, Selbstwahrnehmungen und Vergleiche mit anderen Personen stellen Quellen für selbstbezogenes Wissen und Selbstbewertungen dar (z.B. Gecas, 1982). Externale und internale Referenzen dienen als Rahmen für die Einordnung und Bewertung selbstbezogenen Wissens (z.B. Skaalvik & Skaalvik, 2002). Auf die einzelnen Quellen selbstbezogener Informationen und Referenzrahmen wird im Folgenden detaillierter eingegangen.

1.2.1 Soziale Rückmeldungen und reflektierte Beurteilungen

Erkenntnisse über die eigene Person werden u.a. aus Interaktionen mit der sozialen Umwelt gewonnen (z.B. Gecas, 1982). Nach der Hypothese zum *Looking-Glass-Self* resultieren

Selbsterkenntnisse daraus, dass wir uns in ein Gegenüber hineinversetzen und uns vorstellen, wie wir auf die Person wirken und was die Person über uns denkt (Cooley, 1902). Auch direkte und indirekte Hinweise aus der sozialen Umwelt können Quellen selbstbezogenen Wissens sein. Direkte sprachliche Merkmalszuschreibungen sind z.B. Aussagen wie „Du sprichst gut Deutsch". Indirekt können Menschen aus dem Verhalten anderer darauf schließen, was diese über sie denken (Filipp, 1979; Filipp & Mayer, 2005). Besonders im Schulkontext werden Kinder und Jugendliche kontinuierlich mit direkten und indirekten Leistungsrückmeldungen konfrontiert. Wenn eine Lehrkraft eine Schülerin oder einen Schüler für eine leichte Aufgabe besonders lobt, kann dies z.B. eine *versteckte* Rückmeldung sein (Filipp, 2006). Versteckt ist die Rückmeldung deshalb, weil die Lehrkraft keine explizite Leistungseinschätzung gibt, jedoch ihr besonderes Lob für das Lösen einer einfachen Aufgabe darauf schließen lässt, dass sie der Schülerin oder dem Schüler dies nicht unbedingt zugetraut hätte. Auch zur eigenen Zweitsprachkompetenz kann man immer wieder versteckte Rückmeldungen erhalten, beispielsweise wenn Interaktionspartnerinnen oder -partner nicht verstehen, was man sagt oder wenn diese besonders langsam sprechen.

Verschiedene Befunde sprechen für die Relevanz des sozialen Kontexts als Quelle selbstbezogenen Wissens. Übereinstimmend mit der Hypothese zum Looking-Glass-Self fanden Shrauger und Schoeneman (1979) in einem Review, dass die Selbstwahrnehmung von Personen stark damit zusammenhängt, wie sie sich selbst von anderen wahrgenommen sehen. Einen klaren Zusammenhang zwischen Selbstwahrnehmung und der tatsächlichen Wahrnehmung durch andere konnten sie nicht nachweisen, ebenso wenig konnten sie nachweisen, dass sich Feedback von anderen in natürlichen Situationen auf die Selbsteinschätzungen auswirkt. Methodische Einschränkungen der betrachteten Studien könnten jedoch dafür verantwortlich sein, dass möglicherweise vorhandene Effekte nicht gefunden wurden. Wurde Feedback hingegen experimentell manipuliert, hat sich dies auf die Selbstwahrnehmung ausgewirkt (Shrauger & Schoeneman, 1979). In einer neueren Studie wurde gezeigt, dass der Zusammenhang der Selbstwahrnehmung von deutschen Grundschülerinnen und -schülern mit der Fähigkeitseinschätzung durch Eltern sowie mit der Leistungseinschätzung durch Lehrerinnen und Lehrer im Laufe der Grundschulzeit zunahm. Die Ergebnisse der Studie legen die Vermutung nahe, dass die Einschätzungen der Lehrerinnen und Lehrer sowie die der Eltern eine wichtige Quelle für die Selbsteinschätzungen der Schülerinnen und Schüler darstellen, wobei kausale Analysen ausstehen (Spinath & Spinath, 2005).

Grundsätzlich sind Differenzen zwischen Bewertungen durch andere und Selbsteinschätzungen zu erwarten, da Personen ehrliches Feedback häufig zurückhalten (Blumberg, 1972; Felson, 1980). Höflichkeit stellt ein Hindernis für Feedback als Quelle für Informationen über eigene Eigenschaften oder Kompetenzen dar. Personen versuchen die Gesichter aller Beteiligten zu wahren und keine Gefühle zu verletzen. Dafür enthalten sie negatives Feedback vor, wodurch die Verfügbarkeit von relevanten Informationen eingeschränkt ist (Funder, 1999; Kruger & Dunning, 1999).

Was andere über eine Person denken, ist dieser Person also oft nicht bewusst (vgl. Gecas, 1982). Das heißt, Personen scheinen sich für die Einschätzung durch andere als Quelle für ihre Selbsteinschätzungen zu interessieren, diese Informationen sind jedoch nicht immer zugänglich.

1.2.2 Selbstwahrnehmung

Nach der *Self-Perception-Theorie* von Bem (1972) basieren Selbsteinschätzungen auf der Beobachtung des eigenen Verhaltens. Denn wenn internale Hinweise auf z.B. Emotionen oder Einstellungen schwach, uneindeutig oder nicht interpretierbar sind, müssen Personen auf dieselben Daten und Urteilsprozesse zurückgreifen, um sich selbst einzuschätzen, die sie auch nutzen würden, um eine andere Person einzuschätzen. Dabei geht Bem davon aus, dass internale Hinweise typischerweise schwach und uneindeutig sind. Aus der Selbstbeobachtung wird demnach auf Fähigkeiten und Eigenschaften geschlossen (Bem, 1972). Darüber hinaus werden neben Beobachtungen des eigenen Tuns auch soziale Gruppenzugehörigkeiten für die reflexive Merkmalszuschreibung genutzt (Filipp & Mayer, 2005). Zum Beispiel nimmt man sich als politisch interessiert wahr, wenn man einem Freundeskreis angehört, in dem regelmäßig über politische Themen diskutiert wird.

Wenn auch internale Hinweise nach Bem eine untergeordnete Rolle spielen, können sie dennoch informativ bezüglich eigener Fähigkeiten sein. Solche internalen Hinweise können physiologische Reaktionen, Kognitionen, Emotionen und Motivationen sein (Andersen, 1984; Andersen & Ross, 1984; Bandura, 1977; Markus & Wurf, 1987). Wenn jemand z.B. ungerne Deutsch spricht und dies häufig mit negativen Emotionen wie Frustration einhergeht, schätzt er oder sie seine Deutschkompetenzen wahrscheinlich schlechter ein, als jemand der gerne Deutsch spricht und Spaß daran hat.

Als beispielhaft für die Selbstwahrnehmung können empirische Ergebnisse einer dreiwöchigen Tagebuchstudie interpretiert werden (Niepel et al., 2022). In der Studie wurde u.a. ein Effekt der wahrgenommenen Leistung im Mathematikunterricht auf das mathematische Selbstkonzept festgestellt. Die wahrgenommene Leistung wurde mittels Items gemessen, die abfragten, ob die Schülerin oder der Schüler der letzten Unterrichtsstunde gut folgen konnte, ob sie oder er in der letzten Unterrichtsstunde viel verstanden hat und ob sie oder er in der letzten Unterrichtsstunde viel gelernt hat. Solche Selbstwahrnehmungen scheinen also Quellen selbstbezogenen Wissens zu sein, die das Selbstkonzept speisen. Übertragen auf Sprachkompetenzen sind ähnliche Selbstwahrnehmungen als Quellen des domänenspezifischen Selbstkonzepts denkbar, beispielsweise die Wahrnehmung, ob man gesprochene oder geschriebene Sprache verstehen kann oder wie leicht es einem fällt, sich mündlich oder schriftlich auszudrücken.

1.2.3 Soziale Vergleiche

Nach Festingers (1954) *Social-Comparison-Theorie* hängen die Einschätzungen eigener Fähigkeiten von dem Vergleich mit anderen Personen ab, sofern es kein physikalisches Kriterium gibt, an dem die Fähigkeiten gemessen werden können. Dabei gibt es für die meisten Fähigkeiten kein physikalisches Kriterium, an dem sie gemessen werden können. Und selbst wenn man eine Fähigkeit einfach physikalisch messen kann, wie z.B. die Zeit, die jemand braucht, um eine bestimmte Strecke zu rennen, gewinnt diese Zahl für die Einschätzung der Fähigkeit, schnell zu rennen, an Bedeutung, wenn man sie damit vergleicht, wie viel Zeit andere Personen brauchen, um diese bestimmte Strecke zu rennen (vgl. Festinger, 1954). Nach der Social-Comparison-Theorie sind deshalb für subjektiv genaue Einschätzungen von Fähigkeiten Vergleiche mit anderen Personen notwendig. Dabei vergleichen Personen ihre Fähigkeit bevorzugt mit anderen, die eine ähnliche, aber möglichst bessere Fähigkeit besitzen als sie selbst. Vergleiche werden also bevorzugt nach oben gerichtet (Festinger, 1954; Gerber et al., 2018).

1.2.4 Externale Referenzrahmen

Wenn eine Einordnung eigener Fähigkeiten im Vergleich zu einer Gruppe von Personen stattfindet, spricht man von dieser Gruppe als einem externalen Referenzrahmen. Welche Referenzrahmen für einen Vergleich verfügbar sind und gewählt werden, kann die Einschätzung der Kompetenz beeinflussen und verzerren. Belege dafür liefern das *Shifting-Standards-Modell* (Biernat et al., 1991; Biernat, 2005) und der *Big-Fish-Little-Pond-Effekt* (Marsh, 1987; BFLPE; Marsh & Parker, 1984), welche nachfolgend erläutert werden.

Nach dem Shifting-Standards-Modell kann ein Stereotyp als Referenzrahmen dienen, wenn Mitglieder der stereotypisierten Gruppe hinsichtlich einer für den Stereotyp relevanten Dimension eingeschätzt werden (Biernat et al., 1991; Biernat, 2005). Geht man beispielsweise von dem Stereotyp aus, dass Frauen bessere verbale Fähigkeiten besitzen als Männer, dann wird die verbale Fähigkeit einer Frau mit der angenommenen besseren verbalen Fähigkeit anderer Frauen verglichen, während die verbale Fähigkeit eines Mannes mit der angenommenen schlechteren verbalen Fähigkeit anderer Männer verglichen wird. Der Vergleichsstandard bzw. Referenzrahmen verschiebt sich also je nach Gruppenmitgliedschaft. Folglich könnte bei gleicher verbaler Fähigkeit die verbale Fähigkeit einer Frau niedriger eingeschätzt werden als die eines Mannes, weil unterschiedliche Referenzrahmen angenommen werden und die verbale Fähigkeit der Frau an einem besseren Standard gemessen wird (Biernat, 2005). Als Referenzrahmen werden der Mittelwert und Wertebereich der stereotypisierten Gruppe auf der entsprechenden Dimension angenommen. Im Falle einer Ratingskala nehmen Urteilende an, dass die Punkte die erwartete Verteilung der stereotypisierten Gruppe darstellen und die Endpunkte die minimal und maximal zu erwartenden Werte dieser

Gruppe wiedergeben. Wenn Urteilende einen Unterschied zwischen Gruppen erwarten, dann interpretieren sie auch die Skala und die Endpunkte anders. Dies führt dazu, dass die Einschätzungen nur innerhalb einer Gruppe, nicht jedoch zwischen Gruppen vergleichbar sind, da die Mitglieder verschiedener Gruppen auf unterschiedlich interpretierten Skalen bewertet werden. Dieser Shifting-Standards-Effekt hat sich nur gezeigt, wenn Fähigkeiten oder Eigenschaften auf subjektiven Skalen bewertet wurden, wenn es also möglich war, die Punkte der Skala unterschiedlich zu definieren und anzupassen. Dies ist z.B. bei Likert-Skalen der Fall oder bei Instrumenten mit kontinuierlichen Skalen, die beispielsweise von *sehr niedrige verbale Fähigkeit* bis *sehr hohe verbale Fähigkeit* reichen. Bei objektiven Ratingskalen, die external verankert sind und deren Punkte ihre Bedeutung in verschiedenen Kontexten und bezogen auf verschiedene Individuen nicht verändern, wurde dieser Effekt nicht gefunden. Dies ist z.B. bei der Einschätzung von standardisierten Testscores oder Noten, monetären Einschätzungen, zeitlichen Einschätzungen oder der Erstellung einer Rangordnung der Fall. Bei den objektiven Ratingskalen kommt es hingegen eher zum *Stereotyping-Effekt*, die Einschätzung der Individuen werden an existierende Stereotype angepasst. Die verbale Fähigkeit der Frau wird also tendenziell besser eingeschätzt als die verbale Fähigkeit des Mannes, auch wenn die Fähigkeiten der beiden Individuen tatsächlich gleich sind (z.B. Biernat et al., 1991; Biernat, 2003, 2005; Heine et al., 2002).

Beim BFLPE bildet hingegen kein Stereotyp, sondern das soziale Umfeld den Referenzrahmen für die Einschätzung. Nach dem BFLPE haben Schülerinnen und Schüler mit demselben Kompetenzniveau ein niedrigeres akademisches Selbstkonzept, wenn sie Schulen mit einem höheren durchschnittlichen Leistungsniveau besuchen, als wenn sie Schulen mit einem niedrigeren durchschnittlichen Leistungsniveau besuchen (Marsh, 1987; Marsh & Parker, 1984). Der Effekt scheint darauf zu beruhen, dass sich Schülerinnen und Schüler mit ihrer gesamten Klasse vergleichen, und er existiert neben der Tendenz, dass sich Personen mit anderen mit besserer Kompetenz vergleichen (Huguet et al., 2009). Selbsteinschätzungen hängen also auch davon ab, welche Referenzrahmen in der direkten Umwelt einer Person vorhanden sind, beispielsweise vom Leistungsniveau der Schulklasse.

Zusammenfassend ist anzunehmen, dass die externalen Referenzrahmen individuell unterschiedlich sind und in der Folge die Vergleichbarkeit der Selbsteinschätzungen zwischen Individuen beeinträchtigt ist. Aufgrund des Shifting-Standards-Effekts sind insbesondere die Selbsteinschätzungen zwischen stereotypisierten Gruppen nicht vergleichbar (vgl. Biernat, 2005), vor dem Hintergrund des BFLPE kann jedoch auch die Vergleichbarkeit der selbsteingeschätzten Kompetenz innerhalb einer stereotypisierten Gruppe eingeschränkt sein, sofern diese nicht dasselbe direkte Umfeld teilt (vgl. Marsh & Parker, 1984). Und selbst innerhalb einer Gruppe, die dasselbe direkte Umfeld teilt, können Vergleiche aufgrund unterschiedlicher internaler Referenzrahmen verzerrt sein. Darauf, welche Leistungen und Vorstellungen von Leistungen der eigenen Person als

Referenzrahmen herangezogen werden und wie sich diese auf die Selbsteinschätzungen auswirken, wird im Folgenden eingegangen.

1.2.5 Internale Referenzrahmen

Internal werden verschiedene Vergleichsinformationen zur Selbsteinschätzung herangezogen. Dabei wird zwischen der ideationalen, der temporalen und der dimensionalen Vergleichsebene unterschieden. Auf ideationaler Ebene finden Vergleiche des aktuellen Selbstkonzepts mit Vorstellungen davon, wie man werden könnte, werden möchte, werden sollte oder befürchtet zu werden, statt (Filipp, 1979; Filipp & Mayer, 2005; Skaalvik & Skaalvik, 2002). In diesem Zusammenhang haben Markus und Nurius (1986) den Begriff *Possible Selves* und Higgins (1987) die Begriffe *Ideal Self* und *Ought Self* geprägt, die jeweils eine Referenz für Selbstbewertungen darstellen und mit motivationalen und affektiven Konsequenzen einhergehen (Higgins, 1987; Markus & Nurius, 1986). Auf temporaler Ebene kann eine Kompetenz mit derselben Kompetenz im Zeitverlauf verglichen werden (Albert, 1977). Zuletzt werden, wie in der *Dimensional-Comparison-Theorie* beschrieben (Möller & Marsh, 2013), auf dimensionaler Ebene Leistungen in verschiedenen Domänen miteinander verglichen (Filipp, 2006; Pohlmann et al., 2006; Skaalvik & Skaalvik, 2002). Dimensionale Vergleiche werden insbesondere in bestimmten Situationen ausgelöst, beispielsweise bei der Zeugnisvergabe oder wenn Klassenarbeiten verschiedener Fächer kurz nacheinander zurückgegeben werden (Möller & Köller, 2001). Weiterhin werden dimensionale Vergleiche aus motivationalen Gründen ausgelöst. Dies ist der Fall, wenn Personen Entscheidungen treffen müssen, für die eigene Stärken und Schwächen relevant sind, wie z.B. bei der Kurswahl. Negative Leistungen in einem Fach können dimensionale Vergleiche auslösen, indem der Fokus auf ein Fach gelegt wird, in dem man bessere Fähigkeiten besitzt, um so die Selbstwertschätzung und die Stimmung zu steigern, was eine Form von *Self-Enhancement* (s. Abschnitt 1.3.1) darstellt (Möller & Marsh, 2013). Genauso wie externale Referenzrahmen unterscheiden sich auch internale Referenzrahmen interindividuell, was die interindividuelle Vergleichbarkeit von Selbsteinschätzungen als Kompetenzmaß einschränkt. Eine solche systematische Verzerrung von Selbsteinschätzungen wird im *Internal/External-Frame-of-Reference-Modell* (I/E-Modell; Marsh, 1986) beschrieben.

Gemäß des I/E-Modells beeinträchtigen dimensionale Vergleiche die Korrelation zwischen Selbstkonzept und Fähigkeit in einem Kompetenzbereich. Das mathematische und das verbale Selbstkonzept werden sowohl aus internalen als auch aus externalen Vergleichen abgeleitet. Internal wird das mathematische mit dem verbalen Selbstkonzept verglichen. Dieser internale Vergleich allein würde dazu führen, dass das Selbstkonzept nur in einer der beiden Domänen hoch ist. Damit wäre die Korrelation zwischen mathematischem und verbalem Selbstkonzept negativ. External werden die Kompetenzen mit den Kompetenzen anderer Personen verglichen. Da die mathematische

und verbale Kompetenz in der Regel stark korrelieren, würde der externale Vergleich allein dazu führen, dass das Selbstkonzept in beiden Domänen ähnlich hoch ist. Damit wäre die Korrelation zwischen mathematischem und verbalem Selbstkonzept positiv. Kombiniert man beide Prozesse, erhält man eine Korrelation nahe 0, was auch empirisch so gefunden wurde. Ein hohes mathematisches Selbstkonzept ist demzufolge dann wahrscheinlicher, wenn die mathematische Kompetenz hoch ist und höher ist als die verbale Kompetenz. Der negative Effekt der verbalen Fähigkeit auf das mathematische Selbstkonzept stört also die Korrelation zwischen mathematischer Kompetenz und mathematischem Selbstkonzept (Marsh, 1986). Die Gültigkeit des I/E-Modells wurde metaanalytisch (Möller et al., 2009) und in verschiedenen Stichproben von Schülerinnen und Schülern, in verschiedenen kulturellen Kontexten und in verschiedenen Schulkontexten nachgewiesen (zusammenfassend s. Helm et al., 2020).

Das I/E-Modell wird durch die Dimensional-Comparison-Theorie erweitert, nach der zum einen davon ausgegangen wird, dass dimensionale Vergleiche nicht auf die Erstsprache und Mathematik begrenzt sind und zum anderen, dass dimensionale Vergleiche zwischen ähnlichen Domänen auch zu Assimilations- statt Kontrasteffekten führen könnten. Im Falle von Assimilationseffekten würde sich die Leistung in einem Fach positiv auf das Selbstkonzept in einem anderen, ähnlichen Fach auswirken (Helm et al., 2020; Möller & Marsh, 2013). Das Muster des I/E-Modells wurde auch zwischen anderen Fächern gefunden, beispielsweise zwischen verschiedenen Sprachen (van der Westhuizen et al., 2022). Metaanalytisch hat sich gezeigt, dass der Effekt zwischen Fächern am stärksten war, die auf dem Kontinuum zwischen mathematischen und verbalen Fähigkeiten weit voneinander entfernt liegen. Der Effekt war schwächer aber weiterhin vorhanden, wenn beide Fächer zur verbalen Domäne gehören und nahe null, wenn beide Fächer zur mathematischen Domäne gehören (Möller et al., 2020). Eine mögliche Erklärung dafür, wann Assimilations- und wann Kontrasteffekte auftreten, ist, dass Kontrasteffekte dann auftreten, wenn Personen davon ausgehen, dass die Fähigkeiten, die für die Leistungen in den betreffenden Fächern relevant sind, unterschiedlicher sind, als es tatsächlich der Fall ist. Umgekehrt treten Assimilationseffekte eher dann auf, wenn Personen davon ausgehen, dass die Fähigkeiten, die für die Leistungen in den betreffenden Fächern relevant sind, ähnlicher sind, als es tatsächlich der Fall ist (Helm et al., 2020; Möller et al., 2015). Zusammenfassend sind das I/E-Modell und die Dimensional-Comparison-Theorie ein Beispiel dafür, wie sich internale Referenzrahmen auf Selbsteinschätzungen auswirken.

Die Vielzahl der hier beschriebenen Quellen für selbstbezogene Informationen und unterschiedlichen Referenzrahmen verdeutlicht die Komplexität des kognitiven Prozesses, in dem die verschiedenen Informationen im Selbstkonzept integriert und gewichtet werden (Skaalvik &

Skaalvik, 2002). Weiterhin wird dieser Prozess durch verschiedene Motive beeinflusst, welche im folgenden Kapitel thematisiert werden.

1.3 Motive, die Selbsteinschätzungen beeinflussen

Individuen haben ein besonderes Interesse an ihren Selbsteinschätzungen, denn diese bringen affektive Konsequenzen mit sich (Leary, 2007). Sie streben natürlicherweise nach möglichst angenehmem Affekt (vgl. Gregg et al., 2011). Drei grundlegende Motive, die Selbsteinschätzungen beeinflussen, wurden in der Literatur überwiegend beschrieben: *Self-Enhancement*, *Self-Verification* und *Self-Assessment* (Gregg et al., 2011; Sedikides, 1993; Taylor et al., 1995). Auf die drei Motive sowie auf deren Zusammenspiel wird in den folgenden Abschnitten eingegangen.

1.3.1 Self-Enhancement

Als Self-Enhancement wird die Tendenz zu einem – im Vergleich zur Realität – erhöht positiven Selbstbild bezeichnet, welches mithilfe verschiedener Strategien angestrebt, beibehalten oder erweitert wird (Alicke & Sedikides, 2009; Brown, 1991; Dufner et al., 2019; Taylor & Brown, 1988; Zell et al., 2020). Self-Enhancement manifestiert sich z.B. in Selbsteinschätzungen, die an objektiven Kriterien wie Tests geprüft wurden. Beispielsweise wurden Psychologiestudentinnen und -studenten direkt im Anschluss an eine Prüfung gefragt, wie sie ihre eigene Prüfungsleistung einschätzten, zum einen im Vergleich zu Kommilitoninnen und Kommilitonen und zum anderen bezüglich ihrer erreichten Punktzahl. Insbesondere diejenigen mit den schlechtesten Prüfungsleistungen überschätzten ihre Leistung stark. Die Studentinnen und Studenten im untersten Leistungsquartil schätzten durchschnittlich, dass sie besser als 57% der Prüfungsteilnehmerinnen und -teilnehmer abgeschnitten hätten. Ihre erreichte Punktzahl überschätzten sie um ca. 30% (Dunning et al., 2003).

Darüber hinaus kann Self-Enhancement auch ohne objektives Kriterium nachgewiesen werden. Der prominenteste Effekt, der als Beleg der Self-Enhancement-Theorie herangezogen wird und durch das Self-Enhancement-Motiv erklärt wird, ist der *Better-Than-Average-Effekt* (BTAE; Zell et al., 2020). Der BTAE beschreibt die Tendenz, eigene Fähigkeiten und Eigenschaften positiver einzuschätzen, als die eines durchschnittlichen Peers (Alicke & Govorun, 2005; Zell et al., 2020). Beispielsweise gab die Mehrheit der befragten Autofahrerinnen und Autofahrer an, dass sie besser und sicherer Auto fahren als 50% der anderen Autofahrerinnen und Autofahrer (Koppel et al., 2021; Svenson, 1981). Zahlreiche weitere Studien mit verschiedenen Stichproben, Inhalten und Messinstrumenten bestätigten den BTAE (Alicke & Govorun, 2005). Metaanalytisch erwies sich der BTAE als überaus robust, mit großer bis sehr großer Effektstärke (Zell et al., 2020).

Als adaptiver Nutzen und somit als Begründung des Self-Enhancements wird insbesondere der positive Zusammenhang mit dem psychischen Wohlbefinden hervorgehoben (Dufner et al., 2019; Taylor & Brown, 1988; Zell et al., 2020). Demnach geht Self-Enhancement z.B. mit einem höheren Selbstwertgefühl, einer höheren Lebenszufriedenheit, mehr positivem Affekt, weniger negativem Affekt und weniger Depressivität einher. Inwiefern Self-Enhancement objektive Vorteile mit sich bringt, wie z.B. eine Verbesserung der Leistung, die über eine selbsterfüllende Prophezeiung vermittelt wird, ist umstritten und es wird auch die Möglichkeit von negativen Auswirkungen des Self-Enhancements diskutiert, wie z.B. eine verminderte Anstrengung und somit schlechtere Leistungen als Folge der positiven Illusion oder gefährliche Konsequenzen von Entscheidungen, die auf der Grundlage von falschen Annahmen getroffen wurden (Schütz & Baumeister, 2017). Baumeister (1989) schlug vor, dass positive Verzerrungen des Selbstbilds von geringem bis moderatem Ausmaß optimal sind, da von den affektiven Vorteilen profitiert werden kann, während das Risiko, Entscheidungen mit negativer Auswirkung auf Grundlage falscher Annahmen zu treffen, gering ist.

Abhängig von der zu beurteilenden Fähigkeit oder Eigenschaft neigen Personen mehr oder weniger zum Self-Enhancement. Zwei zentrale dieser moderierenden Faktoren sollen hier genannt werden. Erstens unterliegt dem Self-Enhancement das *Self-Centrality-Breeds-Self-Enhancement-Prinzip* (Gebauer et al., 2013). Nach diesem Prinzip wirkt sich das Self-Enhancement-Motiv insbesondere auf persönlich wichtige Domänen aus (Sedikides & Alicke, 2019). Das Prinzip basiert auf Theorien von William James, welcher folgenden Einblick gab (1890, S. 310): „I, who for the time have staked my all on being a psychologist, am mortified if others know much more psychology than I. But I am contented to wallow in the grossest ignorance of Greek." Auch waren Selbsteinschätzungen umso positiver, je erstrebenswerter die einzuschätzende Eigenschaft im Allgemeinen ist (Alicke, 1985). Zweitens werden Selbsteinschätzungen dann positiv verzerrt, wenn die einzuschätzende Fähigkeit oder Eigenschaft uneindeutiger ist und nicht einfach anhand objektiver Kriterien verifiziert werden kann, wie beispielsweise moralisches Verhalten im Vergleich zu intellektuellen Kompetenzen (Sedikides & Alicke, 2019; van Lange & Sedikides, 1998). Dies wird dadurch erklärt, dass Personen uneindeutige Fähigkeiten oder Eigenschaften eigennützig interpretieren. Beispielsweise kann Führungsqualität daran gemessen werden, inwiefern die Aufgaben und Ziele erfüllt werden oder daran, wie zufrieden die Mitarbeiterinnen und Mitarbeiter sind. Bei der Selbsteinschätzung nehmen Personen bevorzugt die Kriterien an, in denen sie selbst gut abschneiden (Dunning et al., 1989). Je einfacher es bei einer einzuschätzenden Fähigkeit also möglich ist, bevorzugte Kriterien für die Einschätzung stärker zu gewichten, desto größer ist das Ausmaß des Self-Enhancements.

Verschiedene Mechanismen tragen zu einem erhöht positiven Selbstbild bei. Dazu gehören sowohl motivationale Mechanismen, also solche, die aus dem Self-Enhancement-Motiv resultieren,

als auch nicht-motivationale kognitive Mechanismen. Dabei ist die Einordnung in motivational und kognitiv nicht immer eindeutig und für manche Mechanismen gibt es sowohl motivationale als auch nicht-motivationale kognitive Erklärungen (s. Sedikides & Alicke, 2019). Im Folgenden werden einige wichtige Mechanismen erläutert, die zu einer positiven Verzerrung des Selbstbilds beitragen, wobei neben den Mechanismen, die dem Self-Enhancement-Motiv zuzuordnen sind, auch auf Mechanismen eingegangen wird, von denen eher ein kognitiver als ein motivationaler Hintergrund angenommen wird.

1.3.1.1 Selbstwertdienliche Attributionen

Für ein positives Selbstbild nehmen Personen häufig selbstwertdienliche Attributionen vor. Das heißt, sie attribuieren positive Ereignisse internaler, stabiler und globaler als negative Ereignisse (Mezulis et al., 2004). Erfolgen werden internale Ursachen wie eigene Disziplin, Anstrengung und Kompetenz zugeschrieben und Misserfolgen werden externale Ursachen wie die Schwierigkeit der Aufgabe, die Strenge des Gutachters oder Pech zugeschrieben (Sedikides & Alicke, 2019). Auch die Kombination aus internalen jedoch unstabilen und spezifischen Ursachenzuschreibungen für Misserfolge sind möglich, ohne das Selbstwertgefühl zu schädigen, wie z.B. mangelnde Anstrengung (Abramson et al., 1989). Insgesamt werden somit Erfolge häufiger eigenen Fähigkeiten und Eigenschaften zugeschrieben als Misserfolge, was zu einem erhöht positiven Selbstbild beiträgt.

1.3.1.2 Self-Handicapping

Beim Self-Handicapping legen Personen sich selbst Hindernisse in den Weg oder berichten von Hindernissen. Dies erleichtert die Möglichkeit zur selbstwertdienlichen Attribution, denn Misserfolg kann dadurch external und Erfolg internal attribuiert werden (Berglas & Jones, 1978; Schwinger et al., 2014). Relevante Informationen über eigene Fähigkeiten werden so zumindest im Fall von Misserfolg vermieden (Funder, 1999), denn das Hindernis bietet eine Erklärung für schlechte Leistungen und verhindert so, dass diese auf mangelnde Fähigkeiten zurückgeführt werden.

1.3.1.3 Selektives Gedächtnis

Ein weiterer Mechanismus des Self-Enhancement-Motivs ist das selektive Gedächtnis. Personen erinnern sich schlechter an ihre Fehler als an ihre Stärken (Sedikides & Alicke, 2019) und als an die Fehler und Stärken anderer (Sedikides & Green, 2009). Dies gilt für zentrale und wichtige Eigenschaften (Sedikides & Alicke, 2019; Sedikides & Green, 2009). Zu diesem Ergebnis trugen zum einen Studien zum autobiografischen Gedächtnis bei, die zeigten, dass unerfreuliche

Lebensereignisse schlechter erinnert werden als erfreuliche Lebensereignisse (Mather, 2006; Sedikides & Green, 2009). Zum anderen haben experimentelle Labortests mit Gedächtnisabruf-Aufgaben dazu beigetragen. In diesen Studien haben die Teilnehmenden im Anschluss an einen vermeintlichen Persönlichkeitsfragebogen verhaltensbezogenes Feedback zur eigenen Person oder einer anderen fiktiven Person erhalten. Die Teilnehmenden konnten selbstbedrohendes Feedback schlechter abrufen als selbstbestätigendes Feedback oder Feedback, das sich auf andere Personen bezog (Sedikides & Green, 2009). Dieser Effekt wird *Mnemic-Neglect* genannt (Sedikides & Green, 2009). Mnemic-Neglect wird unterdrückt, wenn die Person vorher stärkendes Feedback erhalten hat (Green et al., 2008). Dies stützt den motivationalen Charakter des Effekts (Sedikides & Green, 2009). Weiterhin hat sich der Mnemic-Neglect-Effekt nur für unveränderliche Eigenschaften, nicht jedoch für veränderliche Eigenschaften gezeigt. Dies ist sinnvoll, wenn Eigenschaften mithilfe kritischen Feedbacks verbessert werden können. Eine Regulation des Selbstschutzes scheint also möglich zu sein. Auch dies unterstützt den motivationalen Charakter des Effekts (Green et al., 2005; Sedikides & Green, 2009).

Alternative Erklärungen zum Selbstschutz-Mechanismus (Self-Protection), welcher mit dem Self-Enhancement-Motiv einhergeht, wurden widerlegt. Eine solche alternative Erklärung basiert auf der in Abschnitt 1.3.2 näher beschriebenen Self-Verification-Theorie. Demnach streben Personen nach einem möglichst kohärenten Selbstbild, möchten also ihr Selbstbild erhalten, auch wenn es negativ ist (z.B. Swann & Buhrmester, 2012). Personen mit negativem Selbstbild sollten danach negatives Feedback bevorzugt aufnehmen. Bei Personen mit negativem Selbstbild trat der Mnemic-Neglect-Effekt jedoch ebenso auf (Sedikides & Green, 2004, 2009). Es wird also weiterhin angenommen, dass das selektive Gedächtnis dem Self-Enhancement dient.

Nach der Theorie zum Mnemic-Neglect richten Personen ihre Aufmerksamkeit auf selbstbedrohendes Feedback und enkodieren es, verarbeiten es anschließend aber nur oberflächlich. Dabei wird das negative Feedback separat vom Selbstkonzept verarbeitet und nicht wie positives Feedback in das Selbstkonzept integriert. Insgesamt wird das negative Feedback weniger verarbeitet und kann später schlechter abgerufen werden (Sedikides et al., 2016; Sedikides & Green, 2009).

1.3.1.4 Die Better-Than-Average-Heuristik

Wenn Individuen sich mit durchschnittlichen anderen hinsichtlich einer Eigenschaft vergleichen, nehmen sie automatisch an, dass sie besser sind als diese, ohne überhaupt einen Vergleich vorzunehmen. Diese Annahme wird durch den Befund unterstützt, dass der BTAE auch dann gefunden wird, wenn die Urteile unter kognitiver Beanspruchung stattfinden und somit auf heuristischen Einschätzungen basieren. Die Better-Than-Average-Heuristik basiert auf dem Self-Enhancement-Motiv (Alicke et al., 1995; Beer et al., 2013; Chambers & Windschitl, 2004).

1.3.1.5 Egozentrismus

Ein kognitiver Mechanismus, der eine Selbstüberschätzung begünstigt, wird *Egozentrismus* genannt. Wenn Personen sich mit anderen vergleichen, legen sie den Fokus egozentrisch auf eigene Fähigkeiten bzw. Eigenschaften und berücksichtigen die Fähigkeiten bzw. Eigenschaften der Vergleichsgruppe nur unzureichend. Selbstrelevante Informationen werden im Vergleich zu Informationen über die Vergleichsgruppe im Urteilsprozess also unverhältnismäßig stark gewichtet. Anders ausgedrückt werden bei einfachen Aufgaben die eigenen Vorteile fokussiert, ohne zu berücksichtigen, dass andere womöglich dieselben Vorteile haben, während bei schwierigen Aufgaben die eigenen Nachteile fokussiert werden, ohne zu berücksichtigen, dass andere womöglich dieselben Nachteile haben. Das führt einerseits zum BTAE in allgemein einfachen Aufgaben, wie z.B. eine Computermaus zu bedienen, andererseits zu einem Below-Average-Effekt in allgemein schwierigen Aufgaben, wie z.B. am Computer zu programmieren (Alicke & Govorun, 2005; Chambers & Windschitl, 2004; Kruger, 1999).

1.3.1.6 Focalism

Ein mit dem Egozentrismus verwandter Mechanismus wird in der englischsprachigen Literatur *Focalism* genannt. Breit definiert werden bei Urteilen demnach Aspekte, die im Fokus der Aufmerksamkeit stehen, übermäßig stark gewichtet im Vergleich zu Aspekten, die nicht im Fokus der Aufmerksamkeit stehen (Schkade & Kahneman, 1998). Übertragen auf den BTAE, wird das *Target* einer Selbsteinschätzungsfrage stärker gewichtet, als die Referenz (Chambers & Windschitl, 2004). Dabei ist das Target in der Regel das Selbst, wie in der Frage „Verglichen mit einem typischen Schüler/einer typischen Schülerin, wie viel weißt Du über die Umwelt?“ Die Frage lässt sich jedoch auch so umstellen, dass die ursprüngliche Vergleichsgruppe zum Target wird und somit im Fokus steht: „Verglichen mit Dir, wie viel weiß ein typischer Schüler/eine typische Schülerin über die Umwelt?“ Steht die ursprüngliche Vergleichsgruppe im Fokus, ist der BTAE schwächer, als wenn das Selbst im Fokus der Fragestellung steht (Pahl & Eiser, 2007).

1.3.1.7 Generalized-Group-Account

Eine weitere nicht-motivationale Erklärung des BTAE wird als *Generalized-Group-Account* bezeichnet. Demnach betrachten Personen Individuen positiver als verallgemeinerte Objekte, wie ein durchschnittliches Gruppenmitglied. Personen schätzen also nicht nur sich selbst sondern auch die anderen Individuen einer Gruppe systematisch besser ein als den Gruppendurchschnitt oder Median (Klar, 2002; Zell et al., 2020). Jedoch bleibt der BTAE auch erhalten, wenn sich Personen mit einem bestimmten Individuum vergleichen (Alicke et al., 1995).

Zusammenfassend sind Personen zum Self-Enhancement motiviert und neigen zu einem erhöht positiven Selbstbild (z.B. Alicke & Sedikides, 2009), was sich u.a. im ausführlich untersuchten BTAE äußert (z.B. Zell et al., 2020). Wie stark das Selbstbild in positive Richtung verzerrt wird, hängt z.B. davon ab, wie wichtig einer Person die zu beurteilende Eigenschaft ist und wie eindeutig sie definiert ist bzw. Interpretationsspielraum lässt (z.B. Sedikides & Alicke, 2019). Verschiedene Mechanismen tragen zu dem erhöht positiven Selbstbild bei, wobei manche Mechanismen der Motivation zum Self-Enhancement zuzuschreiben sind und andere kognitiv erklärt werden können (z.B. Alicke & Sedikides, 2009).

1.3.2 Self-Verification

Nach der Self-Verification-Theorie (z.B. Swann, 1983) streben Personen danach, ihr Selbstbild zu erhalten und zu bestätigen, auch wenn das Selbstbild negativ ist (Swann & Buhrmester, 2012). Grundlage der Self-Verification-Theorie ist die Arbeit von Lecky (1945/1969). Dieser versteht das Selbstbild als zentrales, organisiertes Konzept, welches Personen bestrebt sind, beizubehalten. Sobald Personen ein eigenes Selbstbild entwickelt haben, sorgt dieses Selbstbild für ein Gefühl von Kohärenz. Somit erfüllt ein stabiles Selbstbild den Wunsch nach Kohärenz in einer unstabilen Umwelt (Lecky, 1945/1969; Swann & Buhrmester, 2012). Personen sind nach der Self-Verification-Theorie motiviert, das Ausmaß, in dem das Selbstbild durch Erfahrungen bestätigt wird, zu maximieren. Das Streben danach, das eigene Selbstbild durch Self-Verification aufrechtzuerhalten, beginnt in der Kindheit (Cassidy et al., 2003; Swann & Buhrmester, 2012).

Ein stabiles Selbstbild ist deshalb wichtig, weil Personen dieses nutzen, um Vorhersagen über ihre Umwelt zu treffen und das eigene Verhalten zu steuern. Außerdem sind Personen mit einem stabilen Selbstbild berechenbarer für andere, wodurch wiederum die Reaktionen anderer berechenbarer für sie sind und das soziale Umfeld kohärenter ist, was das eigene Selbstbild wiederum bestätigt (Swann & Buhrmester, 2012). Dieser Nutzen wird auch für ein negatives Selbstbild angenommen, denn die Erwartungen anderer und das eigene Verhalten werden gut aufeinander abgestimmt. Problematisch wird das Streben nach der Bestätigung eines negativen Selbstbildes dann, wenn Personen sich ein ungünstiges soziales Umfeld suchen, das sie schlecht behandelt. Insbesondere dieses Phänomen lässt sich mit der Self-Verification-Theorie jedoch erklären (Swann & Buhrmester, 2012). Diese und weitere Strategien, in denen sich das Self-Verification-Motiv ausdrückt, werden als nächstes beschrieben.

Um das eigene Selbstbild zu bestätigen, schaffen sich Personen zum einen eine selbstbestätigende soziale Umwelt. Sie wählen Interaktionspartner, die ein ähnliches Bild von ihnen haben, wie sie selbst, auch wenn das Selbstbild negativ ist. Sie tragen ihre Identität nach außen, z.B. durch

entsprechende Kleidung, um die gewünschten Reaktionen von anderen zu erhalten. Und sie verhalten sich so, dass andere sie so sehen, wie sie sich selbst sehen, auch wenn die andere Person zunächst ein positives Bild hatte und sie selbst ein negatives Bild haben (Swann & Buhrmester, 2012). Zum anderen wenden Personen kognitive Strategien an, um das Selbstbild besser aufrechterhalten zu können (Swann & Buhrmester, 2012). Sie sind empfänglicher für soziales Feedback, wenn sie davon ausgehen, dass dieses ihr Selbstbild bestätigt. Beispielsweise schauen sie sich dieses länger an (Swann & Read, 1981). Außerdem interpretieren Personen Feedback selektiv. Sie nehmen es z.B. nur dann an, wenn es ihrem Selbstbild entspricht (Markus, 1977). Hierbei spielen Attributionen eine Rolle: Personen beziehen Feedback nur dann auf ihre eigenen Eigenschaften, wenn es das Selbstbild bestätigt, ansonsten gehen sie eher davon aus, dass das Feedback mehr über die Quelle des Feedbacks aussagt (Swann et al., 1987). Weiterhin gibt es Hinweise, dass sich Personen an selbstbestätigendes Feedback besser erinnern (Swann & Read, 1981). Dies scheint den oben beschriebenen Befunden zum Mnemic-Neglect zu widersprechen, wonach bei einem negativen Selbstbild trotzdem das positive Feedback besser erinnert wurde (vgl. Abschnitt 1.3.1.3). Auch insgesamt scheinen die Theorien zum Self-Enhancement und zur Self-Verification für Personen mit negativem Selbstbild zunächst widersprüchlich. Es stellt sich also die Frage, welche Theorie gültig ist.

Für die Selbsteinschätzungen von Personen mit positivem Selbstbild werden unter der Annahme des Self-Verification-Motivs dieselben Vorhersagen getroffen, wie unter der Annahme des Self-Enhancement-Motivs, da die Personen unter beiden Annahmen zur Beibehaltung des positiven Selbstbilds motiviert sind. Für die Selbsteinschätzungen von Personen mit negativem Selbstbild unterscheiden sich die Vorhersagen der beiden Motive allerdings, da die Personen unter der Annahme des Self-Enhancement-Motivs dazu motiviert sind, ihr Selbstbild zu verbessern und unter der Annahme des Self-Verification-Motivs dazu motiviert sind, das negative Selbstbild beizubehalten. Als Konsequenz musste nicht eine der beiden Theorien verworfen werden, sondern es hat sich gezeigt, dass die Motive unter verschiedenen Umständen mehr oder weniger dominieren (Kwang & Swann, 2010). Zwei Prinzipien könnten dabei helfen, vorherzusagen, wann welches Motiv dominiert (Swann & Buhrmester, 2012). Nach dem Accessibility-Prinzip benötigen Personen genügend kognitive Ressourcen, um auf das relevante Selbstbild zugreifen zu können, damit eine Bestätigung des Selbstbildes möglich ist (Swann et al., 1990). Demnach ist anzunehmen, dass das Self-Verification-Motiv die Selbsteinschätzungen nur beeinflusst, wenn genügend kognitive Ressourcen zur Verfügung stehen. Self-Enhancement tritt hingegen auch bei eingeschränkten kognitiven Ressourcen auf (Alicke et al., 1995) und beeinflusst die Selbsteinschätzungen, wenn weniger kognitive Ressourcen zur Verfügung stehen. Dafür spricht der Befund, dass Teilnehmende, die die Wahl zwischen einem positiven und einem kritischen Gutachter hatten, unter kognitiver Belastung tendenziell den positiven Gutachter wählten, während Teilnehmende ohne kognitive Belastung

tendenziell den Gutachter wählten, der ihr Selbstbild bestätigte, auch wenn dieser ungünstiger urteilte (Swann et al., 1990). Nach dem Investment-Prinzip sind Personen insbesondere dann motiviert, ihr Selbstbild zu bestätigen, wenn es sich um einen sicheren, fest verankerten Bereich des Selbstbildes handelt. In Bereichen, für die kein sicheres Selbstbild vorliegt, kann auch das Self-Enhancement-Motiv dominieren (Swann & Buhrmester, 2012).

1.3.3 Self-Assessment

Zuletzt gibt es auch Hinweise darauf, dass Personen unter bestimmten Umständen motiviert sind, die eigenen Fähigkeiten möglichst genau einzuschätzen. Insbesondere (in diesem Fall ausschließlich männliche) Studienteilnehmer mit ausgeprägtem Leistungsmotiv haben sich bevorzugt für Aufgaben entschieden, die einen hohen Informationsgehalt bezüglich der eigenen Fähigkeiten hatten (Trope, 1975). Außerdem trat der BTAE eher bei uneindeutigen Fähigkeiten und Eigenschaften auf, nicht jedoch bei eindeutig definierten Fähigkeiten und Eigenschaften (Dunning et al., 1989). Wenn der Interpretationsspielraum also eingeschränkt ist, nehmen Personen genauere Einschätzungen vor (vgl. Gregg et al., 2011). Ebenso war das Self-Enhancement reduziert, wenn Personen davon ausgingen, dass sie ihre Selbsteinschätzungen vor anderen rechtfertigen müssen, dieser Effekt wurde durch die erhöhte Aufmerksamkeit auf Schwächen mediiert (Sedikides et al., 2002). Das Self-Assessment-Motiv scheint also zu dominieren, wenn ansonsten die Glaubwürdigkeit der Selbsteinschätzung gefährdet ist. Darüber hinaus ist eine möglichst genaue Selbsteinschätzung vorteilhaft, wenn sie der Verbesserung der eigenen Fähigkeit oder Eigenschaft dient (Sedikides, 2009) oder wenn es wichtig ist, möglichst genau einzuschätzen, ob man eine bestimmte Herausforderung meistern kann, wie z.B. eine Bucht zu durchschwimmen (vgl. Biernat, 2005; Wheeler et al., 1997). Insgesamt scheint das Self-Assessment-Motiv insbesondere das dominierende Self-Enhancement-Motiv zu ergänzen (vgl. Biernat, 2005) und die beiden Motive halten sich gegenseitig in Schach (vgl. Gregg et al., 2011).

1.4 Entwicklung von Selbsteinschätzungen im Kindes- und Jugendalter

Eigenschaften des Selbstbilds sind abhängig von den kognitiven Strukturen und Einschränkungen des jeweiligen Entwicklungsstands. In der Terminologie von James (1890) beschrieben (vgl. Kapitel 1.1), hängen Struktur und Inhalt des Me-Selbst immer von den dem aktuellen Entwicklungsstand entsprechenden Fähigkeiten des I-Selbst ab. Kognitive Entwicklungen der Prozesse des I-Selbst wirken sich demzufolge auf das jeweilige Selbstbild aus (Harter, 2012). Harter (2006, 2012) beschreibt ausführlich die Entwicklung des Selbst im Kindes- und Jugendalter. Im Folgenden werden zentrale Aspekte ihrer Ausführungen zusammengefasst.

In der sehr frühen bis mittleren Kindheit (im Alter von ca. 2 bis 7 Jahren) sind Selbstbewertungen unrealistisch positiv und die Kinder sind sich der Ungenauigkeit auf naive Weise nicht bewusst. Sie neigen dazu, eigene Leistungen zu überschätzen, z.B., wie weit sie springen können (Schneider, 1998). Die Ungenauigkeiten sind nicht beabsichtigt, sondern resultieren u.a. aus kognitiven Einschränkungen der Entwicklungsstufe. Zum einen fällt es kleinen Kindern schwer, zwischen ihrer gewünschten und tatsächlichen Kompetenz zu unterscheiden, bzw. zwischen idealem und realem Selbstkonzept. Weiterhin sind kleine Kinder noch nicht dazu in der Lage, soziale Vergleiche als Quellen selbstbezogenen Wissens zu nutzen. Temporale Vergleiche können sie hingegen bereits vornehmen, was in der Regel zu positiven Bewertungen führt, da sie die meisten Dinge jetzt besser können als früher. Da sich zumindest die sehr kleinen Kinder noch kaum in andere hineinversetzen können, können sie Bewertungen durch Bezugspersonen noch nicht verstehen und sie nicht als Quellen für selbstbezogenes Wissen nutzen (z.B. Selman, 1980). Schließlich verstehen kleine Kinder auch nicht, dass sie Eigenschaften mit gegensätzlichem Wert besitzen können, also gut und schlecht oder nett und gemein gleichzeitig sein können (Fischer et al., 1984). Für sie ist meistens alles positiv.

Durch das Voranschreiten der kognitiven Entwicklung erlangen Kinder im Alter von ca. 8 bis 10 Jahren ein Bewusstsein sowohl für ihre positiven als auch für ihre negativen Eigenschaften. Weiterhin beginnen sie, soziale Vergleiche für ihre Selbstbewertungen zu nutzen (Ruble & Frey, 1991) sowie reale von idealen Selbstbildern zu unterscheiden und die Bewertungen durch andere Personen zu reflektieren. All das führt dazu, dass die Selbsteinschätzungen der Kinder negativer, aber realistischer werden.

Im frühen Jugendalter (ca. 11 bis 13 Jahre) wird das Selbst in Hinblick auf Domänen und Kontexte zunehmend ausdifferenziert. Mithilfe neuer kognitiver Fähigkeiten können die Jugendlichen einzelne Eigenschaften in Konzepten höherer Ordnung integrieren und Abstraktionen bilden. Die Repräsentationen sind jedoch noch stark voneinander getrennt und die Jugendlichen können die einzelnen Abstraktionen, die das Selbst in verschiedenen Kontexten beschreiben, noch nicht integrieren (z.B. Fischer, 1980; Fischer & Bidell, 2006), weshalb die Selbstbilder inkonsistent und inkohärent sind. Dies stört die Jugendlichen in dem Alter allerdings noch nicht. Insgesamt sind die Selbstbewertungen im frühen Jugendalter sehr ungenau. Die neu gebildeten Abstraktionen lassen sich schwieriger verifizieren und sind häufiger verzerrt. Da die Selbstrepräsentationen noch stark voneinander getrennt sind und immer nur eine abgetrennte Eigenschaft betrachtet werden kann, kommt es zu Übergeneralisierungen, sodass sich die Jugendlichen z.B. einmal für sehr intelligent und einmal für überhaupt nicht intelligent halten. Darüber hinaus beschäftigen sich die Jugendlichen immer mehr mit den reflektierten Beurteilungen durch andere (Cooley, 1902; Harter, 1990; Rosenberg, 1979), wodurch das Selbstbild sehr unbeständig ist, da die Gedanken anderer nicht direkt zugänglich sind und sich die Interpretation ändern kann, da verschiedene

Bezugspersonen verschiedene Meinungen haben und da die Beschäftigung mit der Meinung anderer zu Anstrengungen hinsichtlich der Selbstdarstellung führt, wodurch das Verhalten in verschiedenen Kontexten variiert.

Im mittleren Jugendalter (ca. 14 bis 16 Jahre) entwickelt sich die Fähigkeit, einzelne Abstraktionen bzw. Selbst-Konstrukte miteinander in Beziehung zu setzen (z.B. Fischer, 1980; Fischer & Bidell, 2006). Allerdings können die Jugendlichen diese verschiedenen Selbst-Repräsentationen nicht auf eine Art und Weise integrieren, die Widersprüche löst, wie z.B., dass sie in einer Art von Situation introvertiert und in einer anderen extrovertiert sind. Dass sie sich dieser Widersprüche bewusstwerden, verursacht innere Konflikte und Stress (Harter & Monsour, 1992). Die affektive Komponente, also der Selbstwert ist in diesem Alter tendenziell niedrig. Aufgrund von Verzerrungen kann die Genauigkeit der Selbsteinschätzungen beeinträchtigt sein.

Im späten Jugendalter (ca. 17 bis 19 Jahre) entwickelt sich schließlich die Fähigkeit, Abstraktionen höherer Ordnung zu bilden, sodass einzelne Abstraktionen sinnvoll zusammengeführt werden können (z.B. Fischer, 1980). Somit können Widersprüche reduziert und innere Konflikte gelöst werden (Harter & Monsour, 1992). Es resultiert ein integriertes und kohärentes Selbstbild. Diese Entwicklungsstufe wird jedoch nicht unbedingt automatisch von allen Individuen erreicht. Für das Erreichen der Stufe spielt auch die Unterstützung des sozialen Umfelds eine Rolle, z.B. in Form von der Weitergabe von Erfahrungen oder Erklärungen (Fischer, 1980; Fischer & Bidell, 2006; Karcher & Fischer, 2004). Insgesamt ermöglichen die fortgeschrittenen kognitiven Entwicklungen realistischere Selbstbilder. Weiterhin kennzeichnet diese Stufe, dass die Jugendlichen bzw. jungen Erwachsenen Possible Selves (Markus & Nurius, 1986; vgl. Abschnitt 1.2.5) ins Auge fassen und nach diesen Idealen streben.

Zusammenfassend führen kognitive Limitationen zu stark erhöhten Selbstbildern in der Kindheit. Mit der kognitiven Entwicklung werden Selbstbilder im Verlauf der Kindheit und Jugend realistischer (Harter, 2012).

Aber nicht nur kognitive Limitationen resultieren in erhöhten Selbsteinschätzungen, auch motivational zu erklärende Strategien zum Self-Enhancement werden bereits in der Kindheit angewendet. Metaanalytisch hat sich beispielsweise gezeigt, dass selbstwertdienliche Attributionen von der Kindheit bis ins hohe Alter vorkommen. Während die Verzerrung in der Kindheit sehr stark ausgeprägt ist, nimmt sie in der frühen Jugend drastisch ab und bleibt auf dem abgeschwächten Niveau, bis sie erst im hohen Alter wieder stark ansteigt (Mezulis et al., 2004). Insbesondere mit Blick auf die Kindheit wird dem Self-Enhancement der adaptive Nutzen zugeschrieben, dass es die Motivation und das Engagement erhöht, mit denen verschiedenste neue Aufgaben und Herausforderungen angegangen werden, wie z.B. das Laufen- oder Sprechenlernen. Somit ist ein überhöht positives Selbstbild in der Kindheit förderlich für die Entwicklung (z.B. Trzesniewski et al., 2011). Aber auch später wirkt sich das Selbstkonzept z.B. auf die Leistung in einem Fach aus (z.B. Marsh

& Yeung, 1997). Das heißt, sowohl kognitive Limitationen als auch motivationale Faktoren tragen zu einem verzerrten Selbstbild in der Kindheit und Jugend bei. Bis ins späte Jugendalter wird das Selbstbild zunehmend realistischer und pendelt sich auf einem leicht erhöhten Niveau ein (vgl. Abschnitt 1.3.1).

1.5 Kulturübergreifende Betrachtung der Konstruktion des Selbst

Die bisher beschriebene Literatur stammt überwiegend aus Ländern wie z.B. den USA oder Deutschland, deren Kulturen westlich bzw. individualistisch geprägt sind. Daraus ergibt sich jedoch ein unvollständiges Bild, denn die Konstruktion des Selbst unterliegt kulturellen Einflüssen und unterscheidet sich zwischen individualistisch und kollektivistisch geprägten Kulturen (Markus & Kitayama, 1991). Während sich Personen in individualistisch geprägten Kulturen eher über ihre Einzigartigkeit und Unabhängigkeit von anderen definieren, definieren sich Personen in kollektivistisch geprägten Kulturen mehr über ihre Gruppenzugehörigkeit und streben nach Harmonie innerhalb der Gruppe. Die resultierenden Selbstkonzepte werden als independentes bzw. interdependentes Selbstkonzept bezeichnet (Markus & Kitayama, 1991).

Bei der Konstruktion eines independenten Selbstkonzepts liegt der Fokus auf persönlichen Zielen, Unabhängigkeit und der Abgrenzung von anderen. Personen definieren sich über stabile, kontextunabhängige Eigenschaften, die das Verhalten über verschiedene Situationen hinweg steuern (Cross & Gore, 2012; Markus & Kitayama, 1991). Viele westliche Kulturen sind inzwischen divers und heterogen und Rollenbilder und Erwartungen, die das Verhalten vorschreiben, sind aufgeweicht. Daher ist das Individuum stärker selbst dafür verantwortlich, sich eine Identität zu schaffen, die losgelöster von sozialen Rollen, Status, Geschlecht, Ethnizität und religiöser Erziehung ist. Die Identität wird in modernen westlichen Gesellschaften also persönlich und individuell konstruiert und muss kontinuierlich verifiziert, erneuert und verteidigt werden, auch um eine möglichst kohärente und stabile Identität für sich selbst zu schaffen und diese nach außen tragen zu können. Entsprechend motiviert sind Personen, ihr independentes Selbstkonzept mithilfe der in Kapitel 1.3 beschriebenen Self-Verification- und Self-Enhancement-Strategien zu stärken (Camilleri & Malewska-Peyre, 1997; Cross & Gore, 2012). Für die affektive Komponente, also die Selbstwertschätzung gilt aus US-amerikanischer Perspektive, dass man sich in Bezug auf sich selbst möglichst gut fühlen sollte, auch wenn man dafür selbstrelevante Informationen mithilfe der Self-Enhancement-Strategien verzerren muss (Cross & Gore, 2012).

Bei der Konstruktion eines interdependenten Selbstkonzepts liegt der Fokus auf der Harmonie innerhalb der Gruppe und dem Erreichen von Gruppenzielen (Cross & Gore, 2012; Markus & Kitayama, 1991). In traditionelleren kollektivistisch geprägten Kulturen ist die Identität stärker an

soziale Rollen, das Alter, Geschlecht und den Status gebunden und muss weniger individuell definiert werden. Da das kulturelle System kohärent ist und die Werte von allen akzeptiert werden, wird das Risiko für Widersprüche und internale Konflikte reduziert und die Identität muss kaum verteidigt oder hinterfragt werden (Camilleri & Malewska-Peyre, 1997; Cross & Gore, 2012). In ostasiatischen Kulturen konnten erhöht positive Selbstbilder nur unter bestimmten Umständen nachgewiesen werden, wenn sich die Teilnehmenden mit einem Durchschnitt verglichen, wie z.B. im BTAE, was jedoch auch mit einem nicht-motivationalen kognitiven Mechanismus erklärt werden kann, da Individuen häufig besser eingeschätzt werden als Gruppen (s. a. Abschnitt 1.3.1). Unter anderen Umständen waren Teilnehmende aus ostasiatischen Kulturen eher bescheiden (Heine et al., 2007; Heine & Hamamura, 2007). In asiatischen Kulturen waren selbstwertdienliche Attributionen schwächer ausgeprägt als in den USA oder anderen westlichen Kulturen (Mezulis et al., 2004). Motive zur Selbstverbesserung *(Self-Improvement)* könnten in kollektivistisch geprägten Kulturen stattdessen überwiegen, denn in interdependenten Gesellschaften, in denen die Gruppenzugehörigkeit eine wichtige Rolle spielt, hat es einen hohen Stellenwert, sein Gesicht zu wahren (Ho, 1976) und insbesondere durch andere positiv bewertet zu werden. Dazu ist es wichtig, aufmerksam auf eigene Schwächen zu sein und darauf hinzuarbeiten, diese zu korrigieren (Heine & Hamamura, 2007). Hinsichtlich der Selbstwertschätzung wurden unter Mitgliedern ostasiatischer Gesellschaften niedrigere Ausprägungen festgestellt, als unter Mitgliedern westlicher Gesellschaften, wobei die Selbstwertschätzung mit im Westen entwickelten Skalen gemessen wurde (Cross & Gore, 2012).

Auch im letzten Abschnitt beschränken sich die ergänzenden interkulturellen Betrachtungen zur Konstruktion des Selbst überwiegend auf einige wenige Länder, darunter vor allem die ostasiatischen Länder China und Japan. Für diese kollektivistisch geprägten Kulturen wurden im Vergleich zu individualistisch geprägten Kulturen relevante Unterschiede bei der Konstruktion des Selbst herausgearbeitet (Cross & Gore, 2012; Markus & Kitayama, 1991). In der vorliegenden Arbeit werden größtenteils syrische Geflüchtete betrachtet. Wenn auch die syrische Kultur ganz anders geprägt ist als die ostasiatischen, lassen sich möglicherweise über die Einordnung auf der Kulturdimension des Individualismus und Kollektivismus Implikationen für die Konstruktion des Selbst der syrischen Geflüchteten ableiten. Eine Einordnung von Ländern u.a. auf der Kulturdimension des Individualismus nahm Hofstede (1980, 2001) anhand von Befragungsdaten des Unternehmens IBM aus 72 Ländern aus den Jahren 1967 bis 1973 vor. Die Extremwerte erhielten die USA mit einem Wert von 91 und Guatemala mit einem Wert von 6. Westdeutschland erhielt einen Wert von 67, Hong Kong einen Wert von 25 und Japan einen Wert von 46. Syrien war in die Untersuchung nicht eingeschlossen. Andere arabischsprachige Länder (Ägypten, Irak, Kuwait, Libanon, Libyen, Saudi-Arabien und die Vereinigten Arabischen Emirate) erhielten einen gemeinsamen Wert von 38. Unter der Voraussetzung, dass die syrische Kultur der Kultur der anderen arabischsprachigen

Ländern ähnelt, ist von einer tendenziell kollektivistischen kulturellen Prägung auszugehen, wobei es Hinweise gibt, dass die syrische Kultur zwar immer noch kollektivistischer geprägt ist als z.B. die US-amerikanische Kultur, aber inzwischen individualistischer geprägt ist als früher, was mit der Globalisierung und dem allgemeinen Trend hin zum Individualismus (Stevenson & Zusho, 2002) begründet werden kann (Merkin & Ramadan, 2010). Darüber hinaus ist zu beachten, dass die Geflüchteten seit einiger Zeit in Deutschland leben und sich bereits mehr oder weniger mit den deutschen Normen und Wertvorstellungen identifizieren (Berry, 1997; Esser, 2006). Das heißt, dass sie als Gruppe insgesamt wahrscheinlich individualistischer einzustufen sind, als Personen in ihrem Herkunftsland und dass möglicherweise eine größere interindividuelle Varianz besteht. Wenn auch Implikationen aus der bisherigen Forschung abgeleitet werden können, lassen sich Forschungsergebnisse, die sich auf westliche und ostasiatische Kulturkreise beziehen, nicht einfach auf andere Kulturkreise und auf Personen mit Migrationshintergrund übertragen.

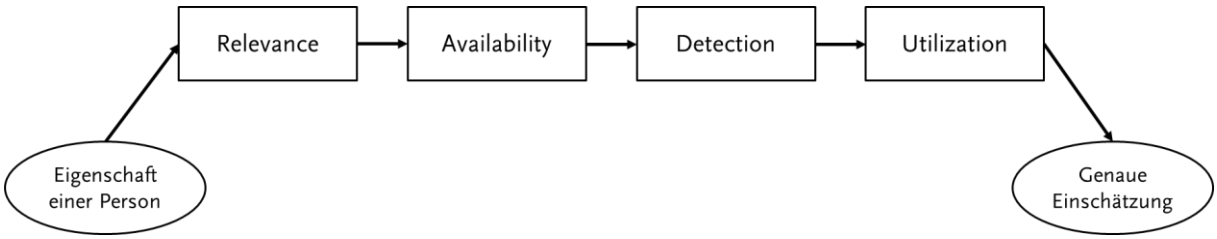
1.6 Das Realistic-Accuracy-Modell (RAM)

Nachdem in den vorangehenden Kapiteln auf verschiedene Aspekte von Selbsteinschätzungen eingegangen wurde, auf deren Zusammenhang mit dem Selbstkonzept, auf verschiedene Quellen selbstbezogener Informationen und Referenzrahmen, auf Motive, die Entwicklung und kulturelle Einflüsse, wird im Folgenden mit dem RAM ein Modell beschrieben, das auf den Prozess der Urteilsbildung eingeht. Einige der oben beschriebenen Aspekte können auf diesem Prozess verortet werden, z.B. Mechanismen des Self-Enhancement-Motivs, die sich an bestimmten Stellen des Urteilsprozesses auswirken. Im RAM beschreibt Funder (1995) die notwendigen Stufen des Prozesses, der von der Eigenschaft einer Person zu einem genauen Urteil über die Eigenschaft der Person führt. Das Modell bezieht sich ursprünglich auf Fremdeinschätzungen von Persönlichkeitseigenschaften und wurde auch auf Selbsteinschätzungen übertragen (Funder, 1999). Es lässt sich neben Persönlichkeitseigenschaften auch auf Kompetenzen anwenden. Aus dem Modell werden Moderatoren der Genauigkeit von Selbsteinschätzungen abgeleitet und systematisiert (Funder, 1995, 1999; Letzring & Funder, 2021). Im Folgenden wird das RAM zunächst kurz beschrieben und es werden Annahmen und Grundlagen des Modells erläutert, bevor in Abschnitt 1.6.1 genauer auf jede der Stufen des RAM sowie auf deren Anwendung auf Selbsteinschätzungen von Sprachkompetenzen und in Abschnitt 1.6.2 auf aus dem Modell abgeleitete Moderatoren der Urteilsgenauigkeit eingegangen wird.

Nach dem RAM ist eine genaue Einschätzung einer Persönlichkeitseigenschaft dann möglich, wenn die folgenden vier Stufen korrekt ablaufen: Zunächst muss die einzuschätzende Person etwas tun, das hinsichtlich der Persönlichkeitseigenschaft relevant ist *(Relevance)*. Zweitens muss

dieses Verhalten so stattfinden, dass es für die urteilende Person sichtbar ist *(Availability)*. Drittens muss die urteilende Person den Hinweis bzw. das Verhalten wahrnehmen *(Detection)*. Zuletzt muss die urteilende Person den Hinweis bzw. das Verhalten in Bezug auf die Persönlichkeitseigenschaft korrekt interpretieren (*Utilization;* Funder, 1995, 2010, S. 205). Die vier Stufen des RAM sind in *Abbildung 1* dargestellt.

Abbildung 1. Das Realistic-Accuracy-Modell (adaptiert nach Funder, 1995, S. 659)



Das RAM beschreibt den vereinfachten Fall, in dem genau eine relevante Information verfügbar ist, wahrgenommen und genutzt wird, um zu einem Urteil über eine einzige Eigenschaft zu kommen. Das Modell beschreibt also den Kernprozess genauer Urteile. In der Realität sind viele relevante Informationen über eine oder mehrere Personen verfügbar und werden wahrgenommen und genutzt, um mehrere Eigenschaften zu beurteilen. Dabei kann eine Information relevant für verschiedene Eigenschaften sein und mehrere Informationen können relevant für dieselbe Eigenschaft sein. In der Realität kann der Prozess von genauen Urteilen also wesentlich komplexer sein, als er im RAM dargestellt wird (Funder, 1995, 1999; Letzring & Funder, 2021). Die vereinfachte Darstellung des RAM ermöglicht dennoch ein besseres Verständnis der Voraussetzungen genauer Urteile sowie die Ableitungen von Moderatoren der Urteilsgenauigkeit (s. Abschnitt 1.6.2), wenn auch die in der Realität höhere Komplexität der Prozesse nicht außer Acht gelassen werden darf.

Das RAM ist ein multiplikatives Modell. Das heißt, wenn eine dieser Stufen nicht erfüllt und somit null ist, gibt es keine genaue Einschätzung. Wenn z.B. keine relevanten Hinweise vorhanden sind, kann eine urteilende Person unmöglich eine genaue Einschätzung treffen. Weiterhin impliziert die Multiplikativität, dass eine perfekte Genauigkeit nur erzielt wird, wenn alle vier Stufen zu 100% erfolgreich sind, wenn also perfekte, eindeutige Hinweise sichtbar sind und optimal wahrgenommen und interpretiert werden. Abgesehen von diesem theoretischen Limit, müssen alle vier Stufen gut erfüllt sein, damit eine substanzielle Genauigkeit erreicht werden kann. Im Modell soll der Erfolg der Stufen jedoch nicht quantifiziert werden, sondern es geht bei der Multiplikativität um die konzeptuelle Idee, dass die Genauigkeit von Selbsteinschätzungen von dem Erfolg jeder Stufe abhängt (Funder, 1995, 1999; Letzring & Funder, 2021).

Eine grundlegende Annahme des RAM ist, dass genaue Einschätzungen (von Persönlichkeitseigenschaften) auftreten können. Diese Annahme ist eine Abgrenzung von der Fokussierung der

Forschung auf Fehler menschlichen Urteilens, wie sie eine Zeit lang vorherrschte (Funder, 1987, 1995, 1999). Dieser auch *Error-Paradigma* genannte (Funder, 1995) Forschungsfokus auf Urteilsfehlern sowie deren zugrundeliegenden Prozesse und Heuristiken (z.B. Ross & Nisbett, 1991) führte zu dem verbreiteten Eindruck, dass menschliche Urteile grundsätzlich ungenau seien. Das Error-Paradigma bot keine Grundlage zur Untersuchung der Genauigkeit menschlichen Urteilens. Erst die Kritik an dem Fokus auf Fehlern und der folgende Paradigmenwechsel hin zur Annahme, dass menschliches Urteilen doch genau sein kann *(Accuracy-Paradigma)*, ermöglichte die Untersuchung der Genauigkeit von Einschätzungen. Denn nur unter der Annahme, dass genaue Urteile möglich sind, ist es beispielsweise möglich zu erforschen, welche Variablen die Genauigkeit von Einschätzungen moderieren (Funder, 1987, 1995, 1999).

Das RAM bezieht sich ausschließlich auf Urteile über reale Personen in realistischen Settings, wenn also eine Realität existiert, über die es zu urteilen gilt. Es bezieht sich also nicht auf Laborsituationen, in denen möglicherweise über fiktive Personen geurteilt wird. Auch dies ist eine Abgrenzung zum Error-Paradigma, denn die Erforschung der Fehler menschlichen Urteilens basierte vorwiegend auf experimentellen Settings, die darauf ausgerichtet waren, Urteilsfehler zu evozieren. Daran wurde kritisiert, dass manche der durch diese Forschung identifizierten Urteilsfehler in der Realität möglicherweise zu korrekten Einschätzungen führen, was jedoch nicht untersucht wurde (Gigerenzer, 1991). Nach dem realistischen Ansatz muss schließlich die Genauigkeit der Urteile im Sinne der Übereinstimmung mit Kriterien der realen Welt überprüft werden (Funder, 1987, 1995; Letzring & Funder, 2021). Darauf, wie die Überprüfung der Genauigkeit von Selbsteinschätzungen erfolgen kann, wird in Kapitel 1.8 eingegangen.

Das RAM lässt sich nicht nur auf Fremdeinschätzungen von Persönlichkeitseigenschaften, sondern auch auf Selbsteinschätzungen von Sprachkompetenzen anwenden. Dass das RAM auf Selbsteinschätzungen übertragen werden kann, begründet Funder (1999, S. 170) mit Bems (1972) Self-Perception-Theorie. Nach der Self-Perception-Theorie basieren Selbsteinschätzungen häufig ebenso auf Verhaltensbeobachtungen, wie es bei Einschätzungen anderer Personen der Fall ist (vgl. Abschnitt 1.2.2). Auf Sprachkompetenzen lässt sich das Modell übertragen, wenn man von einer weiten Definition des Trait-Begriffs ausgeht (vgl. McCrae & Costa, 1995) und die Sprachkompetenz als solchen annimmt. Aus der Übertragung des RAM auf Selbsteinschätzungen von Sprachkompetenzen ergeben sich Besonderheiten für die Anwendung des Modells (Funder, 1999), auf die in der detaillierteren Beschreibung des Modells im folgenden Abschnitt 1.6.1 eingegangen wird.

1.6.1 Die vier Stufen des RAM in Anwendung auf Selbsteinschätzungen von Sprachkompetenzen

Im Folgenden wird auf jede der im vorangegangenen Abschnitt kurz definierten Stufen des RAM sowie auf deren Anwendung im speziellen Fall von Selbsteinschätzungen von Sprachkompetenzen genauer eingegangen.

1.6.1.1 Relevance

Als erste Stufe des RAM muss die Person eine Information freigeben, in der Regel in Form eines Verhaltens, die relevant für den einzuschätzenden Trait ist (Funder, 1995, 1999). Wenn jugendliche Flüchtlinge die deutsche Sprache beispielsweise hauptsächlich passiv, durch Zuhören oder Lesen erwerben, die Sprache jedoch nie sprechen, können sie ihre Kompetenz im Sprechen der deutschen Sprache nicht einschätzen. Dass die jugendlichen Flüchtlinge kein relevantes Verhalten hinsichtlich der deutschen Sprache zeigen, also gar nicht auf Deutsch interagieren, ist vor dem Hintergrund, dass sie in Deutschland leben und eine deutsche Schule besuchen, unwahrscheinlich. Wenn sie hingegen in vielen Kontexten mit verschiedenen Personen auf Deutsch interagieren, z.B. in der Schule, im Sprachkurs und in der Freizeit mit Freunden, ist die erste Stufe des RAM erfüllt und eine genaue Selbsteinschätzung der Deutschkompetenz möglich.

Wie häufig die Jugendlichen relevantes Verhalten zeigen, hängt davon ab, inwiefern sie sich absichtlich in Situationen begeben, in denen sie ihre Deutschkompetenzen testen können oder solche Situationen meiden. Self-Handicapping ist eine Taktik um Gelegenheiten zu vermeiden, bei denen man etwas über eigene Kompetenzen erfahren würde (Funder, 1999). Wie in Abschnitt 1.3.1.2 erläutert, legt man sich beim Self-Handicapping selbst Hindernisse in den Weg oder berichtet von Hindernissen, um Misserfolg external und Erfolg internal attribuieren zu können und so sein Selbstwertgefühl zu schützen (Berglas & Jones, 1978; Schwinger et al., 2014). Um die Überprüfung der eigenen Deutschkompetenzen zu vermeiden oder sich Ausreden zurechtzulegen, könnten die Jugendlichen Interaktionen in deutscher Sprache meiden, indem sie überwiegend Kontakte zu Personen pflegen, mit denen sie ihre Muttersprache sprechen. Manche Personen suchen aber auch absichtlich Situationen auf, in denen sie sich selbst testen können (Funder, 1999). So könnten die jugendlichen Flüchtlinge beispielsweise Kontakt zu Personen suchen, mit denen sie Deutsch sprechen, Medien auf Deutsch nutzen oder an Sprachtests teilnehmen. Die Tendenz zum Self-Handicapping oder Aufsuchen von Situationen, in denen die Deutschkompetenzen getestet werden, kann individuell unterschiedlich sein und sie beeinflusst, wie viel relevantes Verhalten gezeigt wird.

1.6.1.2 Availability

Die zweite Stufe des RAM bezieht sich darauf, dass die relevante Information für die einschätzende Person verfügbar sein muss. Die einschätzende Person muss also beispielsweise anwesend sein, wenn ein relevantes Verhalten ausgeführt wird. Da die einschätzende und die einzuschätzende Person bei der Selbsteinschätzung dieselbe Person ist, ist diese Stufe des RAM hier in der Regel erfüllt und es sind Informationen aus vielen verschiedenen Situationen verfügbar. Darüber hinaus besteht bei der Selbsteinschätzung der Vorteil gegenüber der Fremdeinschätzung, dass auch internale Reize wie unausgesprochene Gedanken und physiologische Reaktionen verfügbar sind (Funder, 1995, 1999; vgl. Mummendey, 2006; vgl. Abschnitt 1.2.2). Weiterhin können Informationen aus sozialen Interaktionen, also soziale Rückmeldungen und reflektierte Beurteilungen, für die Selbsteinschätzung verfügbar sein (vgl. Abschnitt 1.2.1). Direktes Feedback zu eigenen Eigenschaften oder Kompetenzen ist jedoch möglicherweise nur begrenzt verfügbar, wie in Abschnitt 1.2.1 erläutert. Insgesamt sind bei Selbsteinschätzungen also mehr Informationen verfügbar als bei Fremdeinschätzungen, mit möglichen Einschränkungen beim Feedback durch andere.

1.6.1.3 Detection

Auf der Detection-Stufe des RAM geht es darum, ob die einschätzende Person eine relevante und verfügbare Information wahrnimmt, oder nicht. Denn auch wenn eine relevante Information verfügbar ist, registriert man sie möglicherweise nicht. Die Wahrnehmung muss nicht bewusst erfolgen, der Stimulus kann auch unterbewusst wahrgenommen werden und das Urteil unbemerkt beeinflussen. Ob eine relevante und verfügbare Information wahrgenommen wird, kann von verschiedenen Faktoren abhängen, beispielsweise davon, wie aufmerksam man ist und wie salient die Information ist (Funder, 1999).

Beeinträchtigt wird die Aufnahme selbstrelevanter Informationen, wenn man absichtlich wegschaut und so selbstrelevante Informationen meidet (Funder, 1999). Wenn Jugendliche beispielsweise die Korrektur eines Deutschtests nicht anschauen, wäre diese Information über ihre Sprachkompetenz zwar relevant und verfügbar, aber sie würde nicht wahrgenommen und kann nicht zu einer genauen Selbsteinschätzung beitragen. Auch beim Lesen eines deutschen Textes können die Jugendlichen ihre Aufmerksamkeit mehr oder weniger darauf richten, ob sie den Text verstehen.

Manche Personen sind generell aufmerksamer und nehmen selbstbezogene Informationen bewusster wahr (Creed & Funder, 1998). Bei in diesem Sinne Selbst-bewussteren Personen wird die Detection-Stufe vermutlich häufiger erfüllt als bei weniger Selbst-bewussten Personen (Funder, 1999).

Ob eine relevante und verfügbare Information wahrgenommen wird, hängt auch davon ab, ob komplexe Aufgaben bereits einen großen Anteil der kognitiven Ressourcen verbrauchen. Bei

komplexen Aufgaben werden bereits hohe Anforderungen an kognitive Ressourcen gestellt und es bleiben wenige Ressourcen übrig, um die Aufmerksamkeit auf das eigene Verhalten zu richten. Wenn man mit einer anderen Aufgabe beschäftigt ist, während das relevante Verhalten auftritt, besteht die Möglichkeit, dass man dieses Verhalten nicht wahrnimmt, da man seine Aufmerksamkeit auf die andere Aufgabe gerichtet hat (Funder, 1999). Selbst in einfachen sozialen Interaktionen auf Deutsch sind die jugendlichen Flüchtlinge damit beschäftigt, eine andere Sprache zu sprechen, auf die andere Person zu reagieren und sich möglichst angemessen zu verhalten. Hinweise auf die eigene Sprachkompetenz können dabei leicht übersehen werden.

Darüber hinaus kann chronische Ablenkung dazu führen, dass selbstrelevante Informationen nicht wahrgenommen werden. Wenn jemand dauerhaft sehr beschäftigt und gestresst ist, bleiben wenige Ressourcen für die Selbstwahrnehmung. Aber auch psychische Störungen können die Wahrnehmung selbstrelevanter Informationen beeinträchtigen (Funder, 1999).

1.6.1.4 Utilization

Zuletzt muss die urteilende Person die relevante, verfügbare und wahrgenommene Information richtig nutzen (Funder, 1995). Das heißt, sie oder er muss richtig interpretieren, was die Information hinsichtlich ihrer oder seiner Eigenschaft oder Kompetenz bedeutet (Funder, 1999). Dazu gehört auch, richtig zu entscheiden, welche Hinweise relevant für die zu beurteilende Fähigkeit oder Eigenschaft sind, verschiedene Hinweise angemessen zu gewichten und zu kombinieren, und andere Faktoren, die das Verhalten beeinflussen könnten, wie z.B. die Situation, zu berücksichtigen (Letzring & Funder, 2021).

Die richtige Nutzung der Information ist schwierig, weil die Bedeutung eines Verhaltens für eine Eigenschaft oder Kompetenz uneindeutig sein kann. Zum einen spielt der Kontext eine wichtige Rolle bei der Bewertung des Verhaltens (Funder, 1999; Trope, 1986). Wenn beispielsweise eine Jugendliche oder ein Jugendlicher umgangssprachlich mit ihren oder seinen Freunden spricht, ist das hinsichtlich der Deutschkompetenz anders zu bewerten, als wenn die oder der Jugendliche umgangssprachlich im Deutschunterricht oder in einer Deutschprüfung spricht. Zum anderen kann ein Verhalten immer von mehr als einer Eigenschaft oder Kompetenz beeinflusst sein (Ahadi & Diener, 1989; Funder, 1999). Wenn jemandem beim Sprechen vor der Klasse beispielsweise nicht die richtigen Vokabeln einfallen, könnte das auf Nervosität statt auf mangelnde Sprachkompetenz zurückzuführen sein. Weiterhin können Informationen, die über die Eigenschaft oder Fähigkeit bereits vorliegen, die Bewertung wiederum beeinflussen (Trope, 1986), sodass z.B. die fehlenden Vokabeln eher auf mangelnde Sprachkompetenz zurückgeführt werden, wenn bisherige Informationen für eine geringe Sprachkompetenz sprechen und eher auf die Nervosität zurückgeführt werden, wenn bisherige Informationen für eine gute Sprachkompetenz sprechen.

Dabei ist insbesondere die Berücksichtigung von situativen Faktoren bei der Urteilsbildung von den zur Verfügung stehenden kognitiven Ressourcen bzw. der anderweitigen Beanspruchung vorhandener Ressourcen abhängig (Funder, 1999; Gilbert et al., 1988). Unter der kognitiven Belastung des Sprechens vor der Klasse wird die Nervosität aufgrund mangelnder Ressourcen für den Prozess der Urteilsbildung möglicherweise nicht als Ursache für die schlechte Ausdrucksfähigkeit berücksichtigt und das Urteil über die eigene Sprachkompetenz fällt entsprechend negativer aus, als wenn man die Nervosität in das Urteil miteinbeziehen würde.

Die Komplexität der richtigen Interpretation der Informationen bietet auch Spielraum für Verzerrungen. Gerade bei Selbsteinschätzungen kann die Interpretation selbstbezogener Informationen motivational und emotional beeinflusst sein. Darauf, inwiefern Informationen und in der Folge Selbsteinschätzungen tatsächlich verzerrt werden, geht das RAM nicht ein. Das RAM bietet aber eine Grundlage, um den Prozess von Verzerrungen aufzuzeigen (Funder, 1999). Zum Beispiel wurde in Abschnitt 1.3.1.3 beschrieben, dass nach der Theorie zum Mnemic-Neglect Personen selbstbedrohendes Feedback zwar enkodieren, es aber anschließend nur oberflächlich verarbeiten, sodass das negative Feedback separat vom Selbstkonzept verarbeitet und nicht integriert wird. In der Terminologie des RAM bedeutet dies, dass das negative Feedback eine relevante, verfügbare Information ist, die wahrgenommen wird und auch verarbeitet wird. Aber die Verarbeitung auf der letzten Stufe ist unzureichend, um zu einer genauen Selbsteinschätzung beizutragen. Auch selbstwertdienliche Attributionen (vgl. Abschnitt 1.3.1.1) sind auf der Utilization-Stufe des RAM zu verorten. Insbesondere auf der Utilization-Stufe des RAM sind Verzerrungen, u.a. im Sinne des Self-Enhancement zu erwarten, die durch die Komplexität der Aufgabe, die verschiedenen Informationen zu einem Urteil zu integrieren, begünstigt werden.

1.6.2 Moderatoren der Urteilsgenauigkeit

Die Genauigkeit von Einschätzungen kann durch verschiedene Moderatoren beeinflusst werden. Auf einige Moderatoren der Urteilsgenauigkeit lässt sich aus den obigen Ausführungen bereits schließen. Für einen besseren Überblick lassen sich diese in einen Rahmen aus vier übergeordneten Moderatoren einordnen. Die vier übergeordneten Moderatoren sind *Good Judge*, *Good Target*, *Good Trait* und *Good Information* (Funder, 1993). Im optimistischen Sinne sind diese durch das Adjektiv *good* nach der Ausprägung benannt, die eine höhere Genauigkeit ermöglicht.

Die übergeordneten Moderatoren können jeweils miteinander interagieren. Neben den Haupteffekten der einzelnen Moderatoren kann es also auch zu Interaktionseffekten kommen, wenn z.B. eine bestimmte Beziehung zwischen Judge und Trait zu besonders genauen Einschätzungen führt und dies nicht ausschließlich auf die beiden Haupteffekte zurückzuführen ist. Beispielsweise können sich Personen, die sich gut kennen, gegenseitig genauer einschätzen als sich fremde Personen

gegenseitig einschätzen können (z.B. Funder & Colvin, 1988). Dieser Interaktionseffekt zwischen Judge und Target wird als *Relationship* bezeichnet (Funder, 1995). In der Literatur zum RAM wurden die Interaktionen in der Regel separat von den Haupteffekten aufgeführt. Für eine bessere Übersichtlichkeit werden hier jedoch auch Interaktionseffekte bei einer der beteiligten Hauptkomponenten mit aufgeführt.

Die Intention des RAM war, eine theoretische Erklärung genauer Urteile zur Verfügung zu stellen, die allgemein genug ist, um Forschungsergebnisse zu allen Moderatoren zu erklären und weitere Moderatoren vorherzusagen. Das RAM bettet die Moderatoren der Urteilsgenauigkeit also in einen theoretischen Rahmen ein, der zu erklären versucht, warum und wie sich die jeweiligen Moderatoren auswirken. So kann zu jedem Moderator angegeben werden, auf welchen Stufen des RAM er wirkt (Funder, 1999). Die Systematisierung an dieser Stelle dient auch als Überblick und theoretische Einbettung ins RAM für mögliche Moderatoren der Genauigkeit der Selbsteinschätzungen der Sprachkompetenzen jugendlicher Flüchtlinge, die im Rahmen dieser Arbeit untersucht werden sollen.

1.6.2.1 Good Judge

Hinter dem Good-Judge-Moderator steckt die Idee, dass sich Personen darin unterscheiden, wie genau sie andere einschätzen und dass dies mit den Urteilenden selbst zu tun hat (Letzring & Funder, 2021). Gemäß des RAM müssten Unterschiede in der Genauigkeit, mit der verschiedene Personen andere einschätzen, auf die Detection- und die Utilization-Stufe zurückzuführen sein. Es ist also entscheidend, wie gut eine Person verfügbare, relevante Hinweise wahrnehmen und nutzen kann. Wie erfolgreich eine Person darin ist, hängt von drei Komponenten ab: dem Wissen, den Fähigkeiten und der Motivation (Funder, 1999).

Wissen. Das Wissen bezieht sich auf das allgemeine Wissen über Persönlichkeitseigenschaften, bzw. hier übertragen auf das allgemeine Wissen über Kompetenzen und wie sich diese auswirken, z.B. auf das Verhalten. Das Wissen wird durch Erfahrungen im Umgang mit verschiedenen Personen in verschiedenen Kontexten erworben und ist in der Regel implizit. Je mehr Personen über Kompetenzen und deren Manifestation z.B. im Verhalten wissen, desto genauer können sie Kompetenzen einschätzen (Funder, 1999).

In Interaktion mit dem Trait kann sich die *Expertise* der urteilenden Person hinsichtlich des Traits darauf auswirken, wie genau sie dazu in der Lage ist, einen bestimmten Trait einzuschätzen. Im Fall von Kompetenzen ist die Expertise mit der Kompetenz selbst gleichzusetzen. Wenn die urteilende Person nur eine geringe Ausprägung der einzuschätzenden Kompetenz hat, hat sie möglicherweise Schwierigkeiten, die Kompetenz korrekt zu beurteilen und insbesondere Fehler zu

erkennen. Diese von der Ausprägung der Kompetenz abhängige Schwierigkeit bei der Beurteilung der Kompetenz untersuchten Kruger und Dunning (1999) ausführlich.

In einer Serie von Studien fanden Kruger und Dunning (1999) heraus, dass insbesondere Personen, die schlecht in einem Test oder einer Aufgabe abschnitten, ihre Leistung sowohl im Vergleich zu anderen als auch absolut, überschätzten. Die Studien liefen in der Regel so ab, dass Teilnehmende eine Aufgabe, z.B. eine logische Denkaufgabe, bearbeiteten und anschließend ihre logische Denkfähigkeit mit der der anderen Teilnehmenden verglichen. Außerdem schätzten sie ein, wie viele Items sie richtig beantwortet haben. Die Teilnehmenden wurden dann nach ihrer Leistung in Quartile eingeteilt. Diejenigen im untersten Leistungsquartil, die also durchschnittlich den 12. oder 13. Prozentrang erreicht haben, schätzten, dass sie besser als ca. 60% der Teilnehmenden abgeschnitten hätten. Außerdem überschätzten sie die Anzahl der richtig beantworteten Items deutlich. Diejenigen, die am schlechtesten abgeschnitten haben, schätzten ihre Leistung beinahe genauso hoch ein, wie Teilnehmende im oberen Leistungsbereich (Dunning, 2005; Kruger & Dunning, 1999).

Die Autoren argumentierten, dass Personen, die nicht zu einer guten Leistung in der Lage seien, ebenso wenig dazu in der Lage seien, die schlechte Leistung zu erkennen. Sie schrieben von einem *doppelten Fluch*, denn dieselben Fähigkeiten, die notwendig seien, um eine gute Leistung zu erzielen, seien auch notwendig, um zu erkennen, ob man eine gute Leistung erbracht hat. Denn wenn man die Fähigkeit besäße, zu wissen, dass man einen Fehler macht, könne man den Fehler auch vermeiden. Für die kognitive Aufgabe, richtig zu antworten und die metakognitive Aufgabe, die Antwort richtig einzuschätzen, seien dieselben Fähigkeiten notwendig. Dieser Effekt wurde in der Literatur als *Dunning-Kruger-Effekt* bezeichnet (Dunning, 2005; Kruger & Dunning, 1999).

Unterstützt wurde die Argumentation dadurch, dass sich die Genauigkeit der Selbsteinschätzungen auch nicht verbesserte, wenn die Teilnehmenden einen monetären Anreiz erhielten (Ehrlinger et al., 2008). Weiterhin waren die Personen mit besserer Leistung auch besser zu der metakognitiven Aufgabe in der Lage, die Leistung anderer einzuschätzen (Kruger & Dunning, 1999). Darüber hinaus korrigierten Teilnehmende mit zuerst schlechter Leistung in einer Aufgabe ihre Einschätzungen der Leistung, nachdem sie in der Aufgabe effektiv geschult wurden und ihre Fähigkeit verbesserten (Kruger & Dunning, 1999). All diese Befunde stützen die Argumentation, dass es an dem Mangel einer Fähigkeit liegt, weshalb Personen nicht in der Lage sind, ihre Leistung richtig einzuschätzen (vgl. Dunning, 2005).

Jedoch gilt der Dunning-Kruger-Effekt nicht für alle Leistungsbereiche. Personen bemerkten ihre Inkompetenz in der Regel, wenn sie die notwendigen Fähigkeiten überhaupt nicht besaßen (Dunning, 2005). Weiterhin werden in manchen Bereichen unterschiedliche Fähigkeiten zum Erbringen der Leistung und zum Urteilen über die Leistung benötigt. Beispielsweise braucht man andere Fähigkeiten, um die Qualität eines Abschlags im Golf zu beurteilen (z.B. Sehfähigkeit), als

um selbst abzuschlagen (z.B. Koordination; Dunning, 2005). In Bezug auf Zweitsprachfähigkeiten argumentiere ich, dass die Fähigkeit selbst teilweise notwendig ist und teilweise nicht notwendig ist für die Beurteilung, abhängig von der jeweiligen Facette der Sprachkompetenz. Beispielsweise muss man eine Vokabel nicht kennen, um zu bemerken, dass man sie nicht kennt, wenn man sie gerade verwenden möchte. Um den Unterschied zwischen der eigenen Aussprache und der Aussprache von Muttersprachlern zu erkennen, muss man diesen Unterschied hören können. Hingegen muss man die richtige Aussprache hören und die Sprache entsprechend aussprechen können, um akzentfrei zu sprechen. Ist man dazu in der Lage, seine eigene falsche Aussprache zu hören, das Wort aber dennoch nicht richtig auszusprechen, bemerkt man die Inkompetenz wahrscheinlich. In der Grammatik kann man sehr wohl bemerken, dass man z.B. nicht weiß, wie man Vergangenheitsformen bildet. Man kann aber auch falsche Formen bilden oder Formen falsch verstehen und dies mangels Fähigkeit nicht bemerken. Da Sprachkompetenzen verschiedene Fähigkeiten umfassen, sind auch verschiedene metakognitive Fähigkeiten notwendig, um die jeweilige Leistung zu beurteilen. Ich nehme demnach an, dass der Dunning-Kruger-Effekt für Sprachkompetenzen nur in begrenztem Ausmaß gilt.

Eine Schwierigkeit bei der Einschätzung von Kompetenzen sind außerdem die *Errors-of-Omission* (Dunning, 2005, S. 26). Es ist nicht zu greifen, wie groß z.B. der Wortschatz einer Sprache ist und wie viele Möglichkeiten man hätte, etwas sprachlich auszudrücken. Bei Aufgaben, bei denen es, anders als bei z.B. Mathematikaufgaben, nicht nur eine klar definierte Lösung gibt, müsste man theoretisch die gesamte Menge an möglichen Lösungen kennen, um eine Leistung genau einschätzen zu können. Damit geht auch der Befund einher, dass die Selbsteinschätzungen von Schülerinnen und Schülern in naturwissenschaftlichen Fächern besser mit den Einschätzungen ihrer Lehrerinnen und Lehrer übereinstimmten als in sozialwissenschaftlichen Fächern (Dunning, 2005; Falchikov & Boud, 1989).

Die Forschungsarbeiten von Kruger und Dunning (1999) beinhalten noch einen weiteren Befund. Diejenigen Teilnehmenden im oberen Leistungsquartil unterschätzten ihre Leistung beim Vergleich mit den anderen Teilnehmenden. Sie schätzten im Durchschnitt, dass sie nur besser als gut 70% der anderen Teilnehmenden abgeschnitten haben. Nach Kruger und Dunning (1999) lag dies nicht daran, dass sie ihre eigenen Fähigkeiten falsch einschätzten, sondern daran, dass sie die Fähigkeiten der anderen Teilnehmenden überschätzten (vgl. Ehrlinger et al., 2008). Sie nahmen an, dass andere Personen ebenso gut abschnitten und ihre eigene Leistung somit nicht so weit über dem Durchschnitt lag. Diese Argumentation unterstützend, fanden Kruger und Dunning (1999), dass Personen mit sehr guter Leistung ihre relative Selbsteinschätzung erhöht haben, nachdem sie die Tests anderer Personen gesehen haben (vgl. Dunning, 2005).

Fähigkeiten. Weil der Urteilsprozess komplexe kognitive Anforderungen mit sich bringt, könnten sich insbesondere kognitive Fähigkeiten auf die Genauigkeit von Selbsteinschätzungen

auswirken (vgl. Funder, 1999). Für die Anforderungen des Urteilsprozesses scheint insbesondere die kognitive Fähigkeit zum schlussfolgernden Denken relevant. *Fluid Reasoning* wurde definiert als die bewusste und gleichzeitig flexible Aufmerksamkeitskontrolle zur Lösung neuartiger Probleme, die nicht gelöst werden können, indem man ausschließlich bereits erlernte Gewohnheiten, Schemata oder Vorlagen anwendet (Schneider & McGrew, 2012)

Es wird davon ausgegangen, dass Personen mit guten Fähigkeiten im schlussfolgernden Denken mehr kognitive Ressourcen zur Verfügung stehen und sie insgesamt Aufgaben besser lösen können. Bei anspruchsvolleren Aufgaben zeichnen sie sich durch bessere Leistung, bei einfachen Aufgaben durch kürzere Reaktionszeiten aus (van der Meer et al., 2010). Wurde neben der Aufgabenschwierigkeit auch die Art der Aufgaben unterschieden, hat sich gezeigt, dass Individuen mit hoch ausgeprägter Fähigkeit zum schlussfolgernden Denken bestimmte Aufgabenarten effizienter lösen konnten und somit weniger Ressourcen verbrauchten und die Ressourcen in Abhängigkeit von den Anforderungen der Aufgabe flexibler nutzten als Personen mit niedriger ausgeprägter Fähigkeit zum schlussfolgernden Denken (Lu et al., 2022).

Betrachtet man die bisherigen Ausführungen und insbesondere die Utilization-Stufe des RAM, auf der verschiedene Informationen sinnvoll integriert werden müssen, scheint der kognitive Prozess, der zu einer genauen Einschätzung einer Fähigkeit führt, komplex zu sein (s. a.Christiansen et al., 2005). Insbesondere bei Selbsteinschätzungen muss das relevante Verhalten wahrgenommen werden, während es ausgeführt wird (z.B. beim Sprechen muss wahrgenommen werden, ob Vokabeln fehlen oder Fehler gemacht werden), d.h. für die Wahrnehmung und Verarbeitung der relevanten Informationen bleiben nur die kognitiven Ressourcen, die durch das Verhalten selbst noch nicht verbraucht werden. Personen mit hohen kognitiven Fähigkeiten und entsprechenden Ressourcen haben demnach auf der Detection- und Utilization-Stufe des RAM Vorteile, weil ihre kognitiven Ressourcen wahrscheinlicher ausreichen, um trotz bestehender kognitiver Last relevante Informationen wahrzunehmen und verschiedene Informationen sinnvoll zu integrieren und dafür viele Informationen gleichzeitig zu verarbeiten und die Aufmerksamkeit auf die wesentlichen Aspekte zu richten. In Abschnitt 1.7.1 wird darauf eingegangen, dass darüber hinaus in der konkreten Fragebogensituation ebenfalls unter kognitiver Belastung relevante Informationen abgerufen, integriert und in eine Antwort umgesetzt werden müssen. Auch an dieser Stelle spielen kognitive Fähigkeiten eine Rolle für die Genauigkeit der Selbsteinschätzung. Empirisch wurde bisher z.B. zur Fremdeinschätzung von Persönlichkeitseigenschaften gefunden, dass Personen mit höherer Intelligenz bestimmte Persönlichkeitseigenschaften genauer einschätzten (Lippa & Dietz, 2000).

Motivation. Zuletzt ist die Motivation ein entscheidender Faktor auf Seiten der urteilenden Person, der die Urteilsgenauigkeit beeinflussen kann. Demnach wird ein Urteil dann genauer, wenn die urteilende Person zu einer möglichst genauen Einschätzung motiviert ist.

Dementsprechend waren in einer Studie die Übereinstimmungen zwischen Urteilenden größer, wenn die Teilnehmenden davon ausgingen, dass die Urteilsgenauigkeit Konsequenzen für sie hat (Flink & Park, 1991; vgl. Funder, 1999). Bei Selbsteinschätzungen spielen insbesondere die in Kapitel 1.3 beschriebenen Motive eine wichtige Rolle für die Urteilsgenauigkeit. Personen können somit zu einem möglichst positiven Selbstbild (Self-Enhancement), oder zur Bestätigung ihres bestehenden Selbstbilds (Self-Verification) motiviert sein, oder sie können an einer möglichst genauen Selbsteinschätzung interessiert sein (Self-Assessment). Wie stark die einzelnen Motive bei einem Individuum bezüglich einer konkreten Selbsteinschätzung ausgeprägt sind, hängt von verschiedenen Faktoren ab. Auf Seiten der urteilenden Person sind die Stabilität des Selbstbilds, die kulturelle Prägung und die kognitiven Ressourcen und Belastungen relevant. In Interaktion mit dem Trait sind die subjektive Wichtigkeit des Traits und die Eindeutigkeit der zu beurteilenden Kompetenz bzw. Überprüfbarkeit der Selbsteinschätzung relevant (Alicke et al., 1995; Cross & Gore, 2012; Dunning et al., 1989; Gebauer et al., 2013; Mezulis et al., 2004; Sedikides & Alicke, 2019; Swann et al., 1990; Swann & Buhrmester, 2012; van Lange & Sedikides, 1998; vgl. Kapitel 1.3 und 1.5, s. a. Abschnitt 1.6.2.3). Die Motivation der urteilenden Person ist also ein sehr variabler Moderator der Urteilsgenauigkeit und hängt selbst von verschiedenen Faktoren ab.

1.6.2.2 Good Target

Als weiteren übergeordneten Moderator der Urteilsgenauigkeit beschreibt Funder (1993) das Target. Das Target, also die Person, deren Eigenschaft beurteilt wird, ist im Falle der Selbsteinschätzung dieselbe Person wie der Judge, also die oder der Urteilende. Da beim Fokus auf die zu beurteilende Person jedoch andere Eigenschaften auf anderen Stufen des RAM als Moderatoren der Urteilsgenauigkeit wirken, sollen diese auch hier separat von den für die urteilende Person spezifischen Moderatoren betrachtet werden.

Der Hintergrund des Good-Target-Moderators ist, dass sich Personen darin unterscheiden, wie genau sie (nach der ursprünglichen Idee durch andere Personen) bewertet werden. Es scheint an Eigenschaften der Personen selbst zu liegen, dass sie genauer eingeschätzt werden können, denn sie sind selbst dafür verantwortlich, welche Informationen sie preisgeben und für die Aufmerksamkeit, die sie auf sich ziehen (Letzring & Funder, 2021; Mignault & Human, 2021). Im RAM können sich Eigenschaften der einzuschätzenden Person insbesondere auf die Relevanz (Relevance) und die Verfügbarkeit (Availability) von Informationen auswirken. Kompetenzen von Personen, deren Verhalten möglichst viele relevante und unmissverständliche Hinweise preisgibt, sollten am einfachsten einzuschätzen sein (vgl. Funder, 1999). Indirekt können Targets jedoch auch die Detection- und Utilization-Stufe des RAM beeinflussen, indem sie z.B. die Aufmerksamkeit der

urteilenden Person stärker auf sich ziehen oder deren Motivation für eine genauere Einschätzung erhöhen (Mignault & Human, 2021).

Die Menge an verfügbaren Hinweisen kann von dem generellen Aktivitätsniveau einer Person abhängig sein. Manche Personen sind insgesamt aktiver als andere, sodass sie mehr Verhalten zeigen, was probabilistisch gesehen dazu führt, dass mehr relevantes Verhalten verfügbar ist. Für die Einschätzung bestimmter Persönlichkeitseigenschaften spielt insbesondere das Sozialverhalten von Personen eine Rolle. Je schüchterner und sozial zurückgezogener Personen sind, desto weniger Hinweise auf andere Persönlichkeitseigenschaften gibt es, während weniger schüchterne und kontaktfreudigere, extrovertiertere Personen zumindest hinsichtlich mancher Eigenschaften einfacher einzuschätzen sind (Ambady et al., 1995; Funder, 1999; Mignault & Human, 2021). Bei Selbsteinschätzungen von Sprachkompetenzen müssen die Hinweise auf Kompetenzen jedoch nicht zwangsläufig in sozialen Situationen auftreten. Man erhält immer dann Informationen über die eigene Sprachkompetenz, wenn man Kontakt zu der Sprache hat, auch indem man z.B. auf der Sprache schreibt oder liest oder auditive Medien konsumiert. Manche Personen meiden Sprachkontakt und folglich relevante Informationen über eigene Kompetenzen, z.B. als Self-Handicapping-Strategie (s. Abschnitt 1.6.1.1).

In Interaktion mit dem Trait ist die Menge relevanter Informationen, die die einzuschätzende Person verfügbar macht, bei gut sichtbaren Traits jedoch weniger wichtig, denn bei solchen Traits liegen schnell ausreichend Informationen vor, um relativ genaue Einschätzungen vornehmen zu können. Informationen darüber hinaus bieten jedoch wenig Zugewinn für die Genauigkeit der Einschätzung (Funder, 1999). Auch wenn ein Traget also eine geringere Menge relevanter Informationen preisgibt als andere Targets, kann diese Menge bei gut sichtbaren Traits bereits ausreichen, um eine genaue Einschätzung vorzunehmen. Im Fall von Sprachkompetenzen besteht außerdem die Besonderheit, dass durch den Zusammenhang zwischen Sprachkontakt und Sprachkompetenz Personen mit wenig Sprachkontakt auch die Information, dass sie wenig Sprachkontakt haben, nutzen können, um ihre Sprachkompetenz einzuschätzen. Zusammenfassend kann also die Menge relevanter Informationen in einem gewissen Ausmaß durch das Verhalten des Targets beeinflusst werden, was abhängig von der Sichtbarkeit des Traits unterschiedliche Konsequenzen für die Genauigkeit der Einschätzungen haben kann.

Des Weiteren beeinflussen die Konsistenz und Skalierbarkeit der Kompetenz einer Person, wie gut von ihrem Verhalten auf ihre Kompetenz geschlossen werden kann. Es ist möglich, dass die Kompetenz mancher Individuen anders strukturiert ist, als es die Regel ist, dass diese beispielsweise schwierige Vokabeln beherrschen, einige einfache jedoch nicht (vgl. Funder, 1999). Oder dass sie in manchen Bereichen oder Dimensionen der Sprachkompetenz deutlich kompetenter sind als in anderen. Dies kann Schwierigkeiten bei der Einschätzung der Sprachkompetenz bereiten. Ähnlich kann auch ein weiterer z.B. situationsabhängiger Faktor mit der Kompetenz konfundiert sein,

wodurch die gezeigte Kompetenz inkohärent wird (vgl. Abschnitt 1.6.1.1). Wenn jemand z.B. beim Sprechen vor einer Klasse nervös wird und Wörter vergisst und Fehler macht oder wegen Testangst in Sprachtests schlecht abschneidet, während sie oder er in anderen Kontexten deutlich besser in der Sprache interagiert, ist das Verhalten nicht kohärent und entsprechend schwieriger zu beurteilen.

1.6.2.3 Good Trait

Weiterhin beschreibt Funder (1993) den einzuschätzenden Trait als Moderator der Urteilsgenauigkeit. Persönlichkeitseigenschaften oder Kompetenzen können sich darin unterscheiden, wie genau diese tendenziell eingeschätzt werden, d.h. dass manche Traits über verschiedene Urteilende und Beurteilte sowie Situationen hinweg genauer eingeschätzt werden als andere (Letzring & Funder, 2021). Dies hängt einerseits von der Sichtbarkeit der Eigenschaft oder Kompetenz ab, also inwiefern relevantes Verhalten häufig vorkommt und verfügbar ist (Krzyzaniak & Letzring, 2021). Bei Selbsteinschätzungen hängt es andererseits davon ab, wie allgemein erstrebenswert und persönlich wichtig es ist, die Kompetenz zu besitzen, denn das beeinflusst, inwiefern Personen zum Self-Enhancement neigen (Alicke, 1985; Funder, 1999; Sedikides & Alicke, 2019; vgl. Abschnitt 1.3.1). Darüber hinaus beeinflusst die Eindeutigkeit und Verifizierbarkeit anhand objektiver Kriterien einer Eigenschaft oder Kompetenz, wie eigennützig Personen diese interpretieren (Sedikides & Alicke, 2019; van Lange & Sedikides, 1998; vgl. Abschnitt 1.3.1).

1.6.2.4 Good Information

Zuletzt nennt Funder (1993) Informationen als Moderatoren der Urteilsgenauigkeit. Informationen können sich darin unterscheiden, wie nützlich sie sind, um eine genaue Einschätzung vorzunehmen. Es kann also an der Information liegen, ob eine genaue Einschätzung für Urteilende einfach möglich ist. Zwei Aspekte sind dafür entscheidend, wie nützlich Informationen sind: Die Quantität und die Qualität (Letzring & Funder, 2021).

Quantität. Je mehr Informationen man hat, desto genauer wird tendenziell die Einschätzung, die man vornimmt. Dies wird auch mit dem *Aquaintanceship-Effekt* in Verbindung gebracht: Über der oder dem Urteilenden bekannte Personen sind in der Regel mehr Informationen verfügbar, als über fremde Personen, weshalb bekannte Personen genauer eingeschätzt werden können (Funder, 1999; Funder & Colvin, 1988; Letzring & Funder, 2021). Bei Selbsteinschätzungen liegt, wenn man so will, der höchstmögliche Grad an Bekanntheit vor, sodass davon auszugehen ist, dass alle Informationen verfügbar sind (außer Feedback von anderen Personen, vgl. Abschnitt 1.2.1 und Abschnitt 1.6.1.2).

Qualität. Manche Arten von Informationen sind nützlicher, um genaue Einschätzungen vorzunehmen, als andere (Letzring & Funder, 2021). Die Qualität der Information bezieht sich im RAM auf die Relevance-Stufe. Bestimmtes Verhalten in bestimmten Kontexten kann nützlicher sein als anderes Verhalten in anderen Kontexten (Funder, 1999). Die tägliche Begrüßung von Freunden kann weniger informativ über die Deutschkompetenz sein, als im Deutschunterricht einen deutschen Text zu lesen und möglicherweise Lücken auszufüllen. Wenn jemand in spezifischen Situationen relevante Informationen zur eigenen Deutschkompetenz erhält, kann dies zu einer genaueren Selbsteinschätzung beitragen.

Ähnlich wie die Itemschwierigkeit in der Psychometrie die beste Trennschärfe hat, wenn sie der Kompetenz der Person entspricht, sollte auch die *Schwierigkeit* der Situation der Kompetenz der Person entsprechen (Funder, 1999). Wenn eine beginnende Zweitsprachlernerin oder ein beginnender Zweitsprachlerner versucht, sich an einer komplexen politischen Diskussion mit vielen Muttersprachlerinnen und Muttersprachlern zu beteiligen, erfährt sie oder er dadurch vermutlich wenig Neues über ihre oder seine Kompetenz in dieser Sprache, da die Anforderung zu schwierig ist. Aufschlussreicher ist eine einfache Unterhaltung mit einer Freundin oder einem Freund, die dem Niveau der Sprachkompetenz entspricht.

Darüber hinaus kann Feedback von anderen Personen zusätzliche (Quantität), hochwertige (Qualität) Informationen zu einer genauen Einschätzung beitragen (vgl. Bollich et al., 2011). Im Abschnitt zum Good Judge (Abschnitt 1.6.2.1) wurde ausgeführt, dass wenn Personen nicht wissen, was richtig ist, sie auch nicht erkennen können, wenn sie einen Fehler machen (z.B. Dunning, 2005). Genau in solchen Situationen, in denen Personen ihre Fehler nicht wahrnehmen oder entsprechend verarbeiten, kann Feedback besonders hilfreich sein. Dies sollte unabhängig vom Kompetenzniveau einer Person gelten.

1.7 Die Operationalisierung von Selbsteinschätzungen im Fragebogen

Das RAM bezieht sich darauf, wie Urteile gebildet werden, nicht im RAM enthalten ist jedoch der Schritt der Operationalisierung eines Urteils. Die urteilende Person könnte die Einschätzung per Fragebogen oder in einem freien Antwortformat wiedergeben. Die Frage kann einen Referenzrahmen vorgeben oder nicht. Die Genauigkeit des Urteils kann auch in diesem Schritt noch beeinträchtigt werden. Eine Rolle spielt z.B., wie Fragebogenitems formuliert sind und welche Antwortskalen verwendet werden (vgl. Funder, 1995, S. 658, 1999, S. 121). Die kognitiven Prozesse, die bei der Beantwortung von Fragebogenitems ablaufen, sind relevant dafür, wie genau das Urteil wird. Auf diese kognitiven Prozesse wird im folgenden Abschnitt 1.7.1 eingegangen. Implikationen für

die Gestaltung von Selbsteinschätzungsitems, um möglichst genaue Urteile zu erzielen, werden im darauffolgenden Abschnitt 1.7.2 abgeleitet.

1.7.1 Kognitive Prozesse bei der Beantwortung von Fragebogenitems

In einer konkreten Fragebogensituation laufen bei der Beantwortung eines Items folgende Prozesse ab: Das Item muss verstanden werden, die relevante Information muss abgerufen und ein entsprechendes Urteil daraus gebildet werden und schließlich muss eine Antwort auf die Frage gewählt werden (z.B. Strack & Martin, 1987; Tourangeau et al., 2000). Jeder der genannten Prozesse kann aus weiter differenzierbaren Einzelprozessen bestehen und mit Risiken für Ungenauigkeiten einhergehen.

Beim Verständnis der Fragestellung können z.B. verschiedene Teilnehmende Begriffe unterschiedlich verstehen (Tourangeau et al., 2000). Weiterhin kann die Interpretation einer uneindeutigen Fragestellung durch Priming beeinflusst sein, beispielsweise durch vorangegangene Fragen oder andere Aspekte des Kontextes der Befragung (Strack & Martin, 1987). Der Shifting-Standards-Effekt (vgl. Abschnitt 1.2.4) würde möglicherweise nicht auftreten, wenn das Merkmal, aufgrund dessen ein anderer Standard für eine Beurteilung zugrunde gelegt wird, nicht salient wäre.

Manchmal liegen zur Beantwortung einer Frage bereits Informationen im Gedächtnis bzw. im Selbstkonzept vor (Strack & Martin, 1987), die über die im RAM beschriebenen Prozesse gewonnen und zu Urteilen integriert wurden. Wenn die Information zum Zeitpunkt der Beantwortung der Frage verfügbar ist (vgl. Working-Self-Concept; Markus & Wurf, 1987; s. Kapitel 1.1), kann sie einfach genutzt werden. Falls jedoch noch kein integriertes Urteil zu der Fragestellung vorliegt, muss es ad hoc gebildet werden. Das heißt, relevante Informationen müssen aus dem Gedächtnis abgerufen werden, wobei eine vollständige Suche nach Informationen im Gedächtnis in einer Fragebogensituation in der Regel nicht möglich ist. Unter dieser Bedingung entscheidet die Zugänglichkeit darüber, welche Informationen berücksichtigt werden, also beispielsweise, welche Informationen zuletzt aktiviert wurden (Strack & Martin, 1987). Häufig werden in der Fragebogensituation vermutlich sowohl bereits vorliegende Urteile, als auch einzelne Informationen abgerufen und zu einem Urteil integriert (vgl. Tourangeau et al., 2000). Beim Abruf der Informationen ist es möglich, dass die Person den Fokus auf einen bestimmten Aspekt oder einen Teilbereich der Frage legt (vgl. Strack & Martin, 1987). Bei der Frage danach, wie gut jemand Deutsch spricht, könnte die Person beispielsweise eher Informationen und Urteile dazu zusammentragen, wie flüssig, oder wie grammatikalisch fehlerfrei, oder wie akzentfrei, etc. sie oder er die Sprache beherrscht.

Bei der Auswahl einer Antwortkategorie kommt es vor, dass Personen eine Antwortkategorie wählen, die nicht zu ihrem Urteil passt, denn die Bedeutung der Antwortkategorien kann unklar sein und unterschiedlich interpretiert werden oder es könnte keine passende Antwortkategorie

vorliegen (Tourangeau et al., 2000). Auf die Gestaltung von Likert-Skalen und die Auswirkungen auf die Qualität, Reliabilität und Validität der resultierenden Daten wird in Abschnitt 1.7.2 dieses Kapitels eingegangen. Zuvor wird erläutert, unter welchen Bedingungen Teilnehmende die beschriebenen Prozesse bei der Beantwortung von Fragebogenitems so durchführen, dass möglichst genaue Urteile resultieren.

Teilnehmende können aus verschiedenen Gründen motiviert sein, die kognitive Arbeit auf sich zu nehmen, Fragen in einem Fragebogen gründlich und unverzerrt zu beantworten. Häufig liegt jedoch keine Motivation vor, qualitativ hochwertige Antworten zu geben oder die Motivation, die Aufmerksamkeit oder das Interesse lässt im Laufe der Beantwortung eines Fragebogens nach (Krosnick & Presser, 2010). Die Gründlichkeit, mit der Personen Fragen in einem Fragebogen beantworten, beschreibt Krosnick (1991; Krosnick & Presser, 2010) auf einem Kontinuum von *Optimizing* bis *Strong Satisficing*. Beim Optimizing nimmt eine Person den Aufwand auf sich, möglichst optimale Antworten zu geben. Beim Satisficing werden mehr oder weniger extreme Abkürzungen bei den kognitiven Prozessen, die in die Beantwortung der Frage involviert sind, genommen. Das heißt, das Verständnis der Frage, der Abruf aus dem Gedächtnis sowie die Urteilsbildung und die Wahl einer Antwortoption erfolgen weniger gründlich. Im Extremfall werden der Gedächtnisabruf und die Urteilsbildung komplett übersprungen, sodass die Frage oberflächlich interpretiert und eine scheinbar akzeptable Antwort ausgewählt wird. In der Frage wird dann nur nach einem Hinweis gesucht, der auf eine Antwort hindeutet, die einfach auszuwählen und ggf. zu rechtfertigen ist. Dies ist beispielsweise bei der Better-Than-Average-Heuristik der Fall (vgl. Kapitel 1.3).

Wie gründlich Personen bei der Beantwortung von Fragen in einem Fragebogen vorgehen, wird durch drei Faktoren beeinflusst: Durch die Aufgabenschwierigkeit, die Fähigkeiten und die Motivation der teilnehmenden Person (Krosnick, 1991; Krosnick & Presser, 2010). Die Aufgabenschwierigkeit hängt z.B. von Eigenschaften der Frage ab, also wie schwierig die Frage zu verstehen ist und wie schwierig es ist, entsprechende relevante Informationen abzurufen. Auch die Situation der Fragebogenadministration kann Einfluss auf die Aufgabenschwierigkeit nehmen, beispielsweise, ob die oder der Teilnehmende abgelenkt wird. Die Aufgabenschwierigkeit bedingt demzufolge den Umfang der benötigten kognitiven Fähigkeiten und Ressourcen und inwieweit diese z.B. durch Ablenkung, Zeitdruck oder vorangegangene Fragen bereits besetzt sind. Auf Seiten der teilnehmenden Person kommt es darauf an, ob diese dazu in der Lage ist, die komplexen mentalen Operationen unter den gegebenen Bedingungen durchzuführen, wie geübt sie darin ist, über das Thema der Frage nachzudenken und ob bereits entsprechende Urteile oder Einschätzungen vorliegen. Wie motiviert die Person ist, hängt davon ab, ob sie davon ausgeht, dass die Studie sinnvolle Konsequenzen haben wird, ob die Person bereits erschöpft ist, weil die Befragung z.B. schon lange andauert und von der Durchführung der Befragung, also z.B. davon, ob die Interviewerin oder der Interviewer den Eindruck vermittelt, dass die Fragen gründlich oder schnell beantwortet werden

sollen (Krosnick, 1991; Krosnick & Presser, 2010). Außerdem ist es für die Gründlichkeit der Beantwortung der Frage relevant, ob die Person das Thema für persönlich wichtig hält, wobei das im Falle von Selbsteinschätzungen tendenziell zu Self-Enhancement und einer oberflächlichen Beantwortung führt (Gebauer et al., 2013; Sedikides & Alicke, 2019). Fragen zu eigenen Kompetenzen werden eher dann gründlich beantwortet, wenn das Self-Assessment-Motiv überwiegt, was der Fall ist, wenn die Fähigkeit eindeutig definiert ist und wenig Interpretationsspielraum lässt (Dunning et al., 1989; van Lange & Sedikides, 1998) oder die Person davon ausgeht, dass sie ihre Einschätzung später rechtfertigen muss (Sedikides et al., 2002). Neben weiter oben bereits identifizierten Moderatoren der Urteilsgenauigkeit wirken sich folglich auch einige Aspekte der Interviewsituation und des Items im Fragebogen darauf aus, wie gründlich und genau Personen eine Selbsteinschätzungsfrage beantworten.

1.7.2 Implikationen für die Gestaltung von Selbsteinschätzungsitems

Es ist grundsätzlich zu empfehlen, Fragebogenitems möglichst eindeutig und verständlich zu gestalten (z.B. Bühner, 2011). Eine gängige Methode, Selbsteinschätzungen von Fähigkeiten zu erfassen, ist es, eine direkte Frage nach der entsprechenden Fähigkeit auf einer Ratingskala einschätzen zu lassen. Ratingskalen bieten dabei viel Gestaltungsspielraum, denn sowohl die Anzahl der Antwortkategorien als auch deren Labels können variiert werden. Damit alle Teilnehmenden die Antwort auswählen können, die ihrer Einschätzung entspricht, sollte die Antwortskala das gesamte Messkontinuum umfassen und keine Bereiche auslassen (Krosnick & Presser, 2010). Auf einer dichotomen Selbsteinschätzungsskala, die nur die beiden Antwortoptionen *gut* und *schlecht* umfasst, könnten Personen, die ihre Kompetenz als mittelmäßig oder besonders gut oder besonders schlecht einschätzen, beispielsweise keine passende Antwortkategorie wählen. Weiterhin sollten Teilnehmende eine möglichst genaue und stabile Vorstellung davon haben, was die einzelnen Punkte einer Skala bedeuten, sodass sie auch zu einem anderen Zeitpunkt bei gleichbleibender Einschätzung dieselbe Antwort geben würden (Krosnick & Presser, 2010). Darüber hinaus sollten Teilnehmende sowie Forschende in ihrer Interpretation der Bedeutung der Skalenpunkte möglichst übereinstimmen (Krosnick & Presser, 2010), was wenn man die oben beschriebenen Referenzrahmen- und Shifting-Standards-Effekte (vgl. Kapitel 1.2.4 und 1.2.5) bedenkt, häufig nicht gegeben ist.

Insbesondere was die Anzahl der Antwortkategorien und die Labels der Antwortkategorien angeht, gibt es in der Literatur einige empirische Untersuchungen und Empfehlungen. In Bezug auf Meinungsfragen argumentieren Krosnick und Presser (2010), dass fünf Punkte allen Individuen, die neutrale, moderate oder extreme Einstellungen haben, Möglichkeiten geben, diese genau anzugeben. Inwiefern mehr Punkte einen Mehrwert bringen, hängt davon ab, wie präzise die

mentalen Repräsentationen des Konstrukts sind. Mehr Skalenpunkte bringen dann einen Mehrwert, wenn Personen entsprechend präzisere Unterscheidungen vornehmen und tatsächlich die gesamte Skala nutzen. Wird die Skala länger, müssen die Teilnehmenden die Bedeutung von mehr Punkten interpretieren und es besteht die Gefahr, dass sie die Punkte weniger konsistent interpretieren. Darüber hinaus könnte eine hohe Anzahl an Punkten die Aufgabenschwierigkeit erhöhen, da mehr Punkte interpretiert und zwischen mehr Punkten gewählt werden muss. Andererseits könnten zu wenige Punkte die Aufgabenschwierigkeit ebenfalls erhöhen, wenn dadurch keine passende Kategorie vorhanden ist. Entsprechend empfehlen Krosnick und Presser (2010), Antwortskalen moderater Länge zu verwenden. Ihre Zusammenfassung empirischer Befunde unterstützt die Empfehlung. Sie ergab, dass die Reliabilität von Skalen mit nur zwei oder drei Punkten geringer ist als die Reliabilität von Skalen mit mehr Punkten, aber dass die Reliabilität bei mehr als 7 Punkten nicht weiter ansteigt. Bei der Validität zeichnet sich ein ähnliches Bild. Leung (2011) hat hingegen die Verteilung von Antwortskalen mit 4, 5, 6 und 11 Optionen anhand einer chinesischen Version der Selbstwertskala nach Rosenberg (Leung & Wong, 2008) untersucht, mit dem Ergebnis, dass die 11-stufige Skala eine geringere Schiefe und Kurtosis hatte und einer Normalverteilung am nächsten kam, wobei sich die 11-stufige Skala verglichen mit den anderen Skalen nicht auf anderen Kennzahlen wie dem Mittelwert, der Standardabweichung, der Reliabilität und Validität unterschied. Leung (2011) empfiehlt den Gebrauch einer 11-stufigen Skala von 0 bis 10, mit der Argumentation, dass so die Empfindlichkeit der Skala erhöht wird und es jederzeit möglich ist, nachträglich Punkte zu gemeinsamen Kategorien zusammenzulegen. Das Problem, dass zu viele Punkte die Aufgabenschwierigkeit erhöhen, schätzt er als gering ein, da eine Skala von 0 bis 10 sehr geläufig und einfach zu verstehen sei. Die Empfehlungen zur Anzahl der Antwortkategorien liegen also im Bereich von 5 bis 11 Kategorien, wobei verschiedene Argumente für mehr oder weniger Kategorien sprechen.

Die Frage nach Labels für die Antwortkategorien ist in Zusammenhang mit der Anzahl der Kategorien zu betrachten, denn je mehr Antwortkategorien eine Skala umfasst, desto schwieriger wird es, jede einzelne Kategorie mit einem Label zu versehen (vgl. Krosnick & Presser, 2010). Je nach intendierter Nutzung der Daten soll die Ratingskala möglichst an eine Intervallskala angenähert sein. Das suggeriert in der Regel auch die Darstellung der Ratingskala im Fragebogen durch konstante Abstände zwischen den Punkten oder z.B. als Schieberegler. Grundsätzlich helfen Labels dabei, die Bedeutung jedes Punktes der Skala zu definieren. Jedoch kann auch die Bedeutung von Labels durch verschiedene Personen unterschiedlich interpretiert werden. Darüber hinaus können die Labels die Interpretation der Skala als Intervallskala, bei der die Abstände zwischen den Kategorien gleich groß sein sollen, beeinträchtigen. Dies ist der Fall, wenn die Abstände zwischen den Labels nicht gleich interpretiert werden. Eine gängige Alternative ist es, nur die Endpunkte der Skala mit einem Label zu versehen (Cummins & Gullone, 2000; Leung, 2011). Das heißt, auch zu

den Labels gibt es keine eindeutige Empfehlung, sondern verschiedene Argumente sprechen für verschiedene Möglichkeiten.

Nachdem die Befunde und Implikationen zur Gestaltung der Antwortskala dargelegt wurden, bleibt es, auf die Formulierung der Selbsteinschätzungsfrage einzugehen. Zum einen sollte die Frage möglichst einfach und verständlich formuliert sein, um so die Aufgabenschwierigkeit zu reduzieren und die Gründlichkeit bei der Beantwortung zu fördern (vgl. Abschnitt 1.7.1). Weiterhin sollte die Frage so eindeutig formuliert sein, dass sie wenig Interpretationsspielraum bietet und Teilnehmende nicht unterschiedliche Kompetenzbereiche bewerten (vgl. Dunning et al., 1989; Kapitel 1.3.1). Zum anderen konnte metaanalytisch gezeigt werden, dass relative Antwortskalen und Instruktionen, die auf einen Vergleich mit anderen hinwiesen, bzw. die Vorgabe einer Referenzgruppe, die Validität von Selbsteinschätzungen erhöhen (Freund & Kasten, 2012; Mabe & West, 1982). Dies ist einleuchtend vor dem Hintergrund, dass Personen unterschiedliche Referenzrahmen bei der Beantwortung von Selbsteinschätzungsfragen annehmen (vgl. Abschnitte 1.2.4 und 1.2.5 zu externalen und internalen Referenzrahmen), was zu einer unterschiedlichen Interpretation der Antwortskala führt. Ist der Referenzrahmen vorgegeben, könnte die Interpretation der Antwortskala angeglichen werden. Zuletzt tragen auch die Vertrautheit und Erfahrung mit der Fähigkeit zu genaueren Selbsteinschätzungen bei (Freund & Kasten, 2012). Selbsteinschätzungen waren beispielsweise genauer, wenn im Unterricht Erfahrungen mit der einzuschätzenden Fähigkeit gemacht wurden (Ross, 1998) oder wenn sich die Selbsteinschätzungsfragen sehr konkret auf die Situation von Teilnehmenden als potenzielle Zweitsprachnutzer bezogen und z.B. durch eine Aussage der folgenden Art eingeleitet wurden: „Wenn ich mich mit einer französischsprachigen Studentin unterhielte, wäre ich in der Lage, folgendes zu verstehen (...)“ (LeBlanc & Painchaud, 1985). Für Teilnehmende scheint es also einfacher zu sein, Selbsteinschätzungsfragen zu beantworten, die sich auf bekannte Situationen beziehen. Zusammenfassend sind Fragen idealerweise einfach und eindeutig formuliert, geben einen Referenzrahmen vor und beziehen sich auf vertraute Fähigkeiten oder Situationen. Diese Empfehlungen bedürfen jedoch dringend (weiterer) empirischer Prüfung.

1.8 Erfassung der Genauigkeit von Selbsteinschätzungen

In dieser Arbeit geht es vordergründig um die Untersuchung der Genauigkeit von Selbsteinschätzungen. Deshalb spielt auch die Erfassung der Genauigkeit von Selbsteinschätzungen eine wichtige Rolle. Im Folgenden wird auf Kriterien zur Erfassung der Genauigkeit von Selbsteinschätzungen eingegangen und auf die Relevanz der Übereinstimmung der Domänen, die mit der Selbsteinschätzung und den Kriterien erfasst werden. Darüber hinaus werden verschiedene Indikatoren

erläutert, die zur Messung verschiedener Komponenten der Genauigkeit von Selbsteinschätzungen dienen.

Die Überprüfung der Genauigkeit der Urteile kann wie eine Testvalidierung erfolgen (Funder, 2010, 190f). Zur konvergenten Validierung können verschiedene Kriterien herangezogen werden. Dazu gehören beispielsweise die Übereinstimmung zwischen verschiedenen Urteilenden, standardisierte Testverfahren, Schulnoten und Verhaltensbeobachtungen (z.B. im Labor). Darüber hinaus ist die prädiktive Validität der Urteile relevant, also inwiefern das Urteil zukünftiges Verhalten oder bestimmte Outcomes vorhersagen kann. Funder (1995; vgl. auch McCrae, 1982) empfiehlt, möglichst viele verschiedene Kriterien zu verwenden, denn jedes Kriterium hat Vor- und Nachteile. Zum Beispiel können – wie im Falle von ethnischen Vorurteilen – viele verschiedene Urteilende zu demselben Ergebnis kommen, aber alle falsch liegen. Werden zusätzliche andere Kriterien verwendet, könnte das Ergebnis ein anderes sein, sodass der Fehler auffällt. Auch wenn es nicht möglich ist, alle gewünschten Kriterien zu sammeln, sollte es nach dem RAM das Ziel sein, möglichst viele Kriterien in eine Studie einzuschließen (Funder, 1995, 2010, 190f).

Studien zum BTAE (vgl. Abschnitt 1.3.1) kamen auch ohne die Hinzunahme eines Kriteriums aus. In diesem Fall wurde die Tatsache, dass sich die meisten Teilnehmenden als besser als eine durchschnittliche Person einschätzten als Beleg einer allgemeinen Tendenz zum Self-Enhancement auf Gruppenebene angenommen. Dieses Vorgehen hat jedoch den Nachteil, dass interindividuelle Differenzen in der Neigung zum Self-Enhancement nicht erfasst werden, da auf individueller Ebene nicht bekannt ist, ob eine Über- oder Unterschätzung vorliegt. Sofern also interindividuelle Unterschiede in der Genauigkeit der Selbsteinschätzung von Interesse sind, ist es notwendig, ein Außenkriterium zu erheben (vgl. Schütz et al., 2016).

Weiterhin müssen, um die Genauigkeit von Selbsteinschätzungen beurteilen zu können, die Domänen, die mit der Selbsteinschätzung und dem Kriterium erfasst werden, möglichst übereinstimmen. Sprachkompetenzen sind facettenreich und mehrdimensional. Sie umfassen verschiedene Arten von Wissen und Fähigkeiten, u.a. grammatisches Wissen und Wortschatz, welche für die Produktion und das Verständnis von geschriebener und gesprochener Sprache relevant sind. Die Sprachkompetenz von Individuen kann zwischen verschiedenen Domänen variieren. Individuen können z.B. einen guten Wortschatz haben, aber grammatikalische Fehler machen. Selbsteinschätzungen werden häufig durch eher globale Fragen wie “Wie gut sprechen Sie Deutsch?“ erfasst. Solche globalen Fragen definieren nicht die Domäne oder den Aspekt, welcher beurteilt werden soll. Folglich kann sich die Domäne, die die einschätzende Person annimmt, von der Domäne unterscheiden, die das Kriterium (z.B. der Test) misst (vgl. Artelt, 2016; Gollan et al., 2012). Bezüglich der Beurteilung von Schülerinnen- und Schülerleistungen durch Lehrerinnen und Lehrer argumentiert Artelt (2016), dass die Übereinstimmung zwischen der beurteilten Domäne und dem gemessenen Kriterium die Einschätzung der Urteilsgenauigkeit moderiert. Sie bezieht sich

auf eine Metaanalyse, die zu diesem Ergebnis kam: In Studien, in denen die Spezifizierung der Domäne inkongruent war, wurde die Urteilsgenauigkeit niedriger eingeschätzt, als wenn diese kongruent war (Südkamp et al., 2012). Wenn sich die Selbsteinschätzung also auf allgemeine deutsche Sprachkompetenzen bezieht, jedoch nur der rezeptive Wortschatz gemessen wird, könnte dies die Korrelation zwischen der Selbsteinschätzung und dem Kriterium beeinträchtigen und zu einer Unterschätzung der Genauigkeit der Selbsteinschätzungen führen.

Bisher wurde argumentiert, dass Kriterien zur Überprüfung der Genauigkeit von Urteilen herangezogen werden und diese möglichst mit der durch die Einschätzung erfassten Kompetenz übereinstimmen sollten. Als nächstes wird darauf eingegangen, auf welche verschiedenen Weisen die Selbsteinschätzung mit dem Kriterium verglichen und welche Indikatoren dazu gebildet werden können. Verschiedene sich ergänzende Indikatoren können die Urteilsgenauigkeit messen, denn die Genauigkeit von Selbsteinschätzungen kann als Kombination aus mehreren Komponenten betrachtet werden. Eine Komponente ist die Diskrimination, also inwiefern die Selbsteinschätzungen die korrekte Rangordnung innerhalb der untersuchten Gruppe wiedergeben können. Diese Komponente kann durch Korrelationskoeffizienten quantifiziert werden: Je höher die Korrelation zwischen Selbsteinschätzung und Kriterium, desto genauer sind die Selbsteinschätzungen. Dieser Indikator der Urteilsgenauigkeit wird in psychologischen Studien vorwiegend verwendet (Dunning, 2005). Eine weitere Komponente ist die Verzerrung, also der allgemeine Fehler, den Personen bei der Selbsteinschätzung ihrer Leistung oder Kompetenz machen (Epley & Dunning, 2006). Die beiden Komponenten sind statistisch unabhängig, sodass auch bei starker Korrelation eine deutliche Verzerrung (z.B. allgemeine Überschätzung der Kompetenz) vorliegen kann (vgl. Artelt, 2016; Funder & Colvin, 1997; John & Robins, 1994). Eine weitere Komponente der Urteilsgenauigkeit ist die Variation der Werte innerhalb der Gruppe. Auch wenn die Rangordnung und das allgemeine Level der Fähigkeiten durch die Selbsteinschätzungen richtig abgebildet werden, kann unabhängig davon die Variation der Werte in der Gruppe durch die Selbsteinschätzungen nicht richtig erfasst werden (vgl. Artelt, 2016). Zusammenfassend stellt das Korrelationsmaß ein wichtiges, nicht jedoch hinreichendes Maß der Genauigkeit von Selbsteinschätzungen dar. Es ist ebenso wichtig, die Verzerrungen von Selbsteinschätzungen zu messen, da sonst allgemeine Über- oder Unterschätzungen übersehen werden (vgl. Dunning & Helzer, 2014). Dennoch wurden in Studien zur Genauigkeit von Selbsteinschätzungen überwiegend Korrelationskoeffizienten berechnet, wie im folgenden Kapitel deutlich wird.

1.9 Empirische Befunde zur Genauigkeit von Selbsteinschätzungen

In diesem Kapitel werden Studien zur Validität von Selbsteinschätzungen zusammengefasst. Zur Validität von Selbsteinschätzungen gibt es zahlreiche Studien und Metaanalysen. Bei einem Großteil der Studien wird nicht die Validität von Selbsteinschätzungen als Sprachkompetenzmaß untersucht, sondern die Validität der Selbsteinschätzungen verschiedener Kompetenzen. Obwohl Unterschiede zwischen der Validität der Selbsteinschätzungen verschiedener Kompetenzen möglich sind, so sind sie vermutlich nicht grundlegend verschieden und die Ergebnisse solcher Studien könnten auch hinsichtlich der Genauigkeit der Selbsteinschätzungen von Sprachkompetenzen aufschlussreich sein.

Die Validität, die hier betrachtet wird, ist die Konstruktvalidität von Selbsteinschätzungen als Maß für Sprachkompetenzen, d.h. es geht darum, ob Selbsteinschätzungen die tatsächliche Sprachkompetenz erfassen, was über den Zusammenhang mit anderen Sprachkompetenzmaßen untersucht wird. Davon abzugrenzen ist die Validität von Selbsteinschätzungen von Sprachkompetenzen, welche zum Inhalt hat, inwiefern Selbsteinschätzungen erfassen, wie die Personen tatsächlich ihre Sprachkompetenzen einschätzen, also inwiefern die Selbsteinschätzungen das Selbstkonzept der entsprechenden Sprachkompetenzen wiedergeben. Selbsteinschätzungen können das Selbstkonzept der Sprachkompetenz valide erfassen und sich dennoch deutlich von der tatsächlichen Sprachkompetenz unterscheiden, wenn das Selbstkonzept von der objektiv gemessenen Sprachkompetenz abweicht. Die Validität von Selbsteinschätzungen als Maß für Sprachkompetenzen steht hier im Fokus, da die Selbsteinschätzungen als alternatives Kompetenzmaß in Studien angewendet werden und sich die Frage stellt, inwiefern das die Ergebnisse von Analysen zu Sprachkompetenzen beeinflusst. Inwiefern Selbsteinschätzungen das Selbstkonzept der Sprachkompetenz wiedergeben, wäre eine in anderer Hinsicht relevante Fragestellung, denn das Selbstkonzept kann u.a. Entscheidungen, wie z.B. Bildungsentscheidungen, beeinflussen (z.B. Guo et al., 2015). Dies steht jedoch nicht im Fokus dieser Arbeit.

In Metaanalysen zur Validität von Selbsteinschätzungen verschiedener Kompetenzen wurden Korrelationen mittlerer Effektstärke zwischen Selbsteinschätzungen und Kriterien gefunden. Mabe und West (1982) haben 55 Studien analysiert, in denen Selbsteinschätzungen von Kompetenzen mit objektiven Tests, Noten und Vorgesetzteneinschätzungen verglichen wurden. Unter Berücksichtigung von 267 Korrelationskoeffizienten fanden sie eine mittlere Korrelation von $r = .29$ ($SD = .25$). In einer weiteren Metaanalyse, in die Studien eingeschlossen wurden, in denen Selbsteinschätzungen mit Einschätzungen von Lehrerinnen und Lehrern verglichen wurden, wurde unter Berücksichtigung von 46 Korrelationskoeffizienten eine mittlere Korrelation von $r = .39$ gefunden (Falchikov & Boud, 1989). In einer jüngeren Metaanalyse zur Validität von

Selbsteinschätzungen kognitiver Fähigkeiten wurde eine mittlere Effektstärke von $r = .33$ gefunden (Freund & Kasten, 2012). In dieser Metaanalyse wurden 154 Korrelationen zwischen Selbsteinschätzungen und Testscores aus 41 Studien berücksichtigt.

Um den Zusammenhang zwischen Selbsteinschätzungen von Sprachkompetenzen und einem Kriterium geht es in nur einer mir bekannten Metaanalyse. In dieser Metaanalyse lag der Mittelwert aus 60 Korrelationskoeffizienten aus 10 Studien bei $r = .63$ mit einer großen Varianz (Ross, 1998). Dieser Wert ist deutlich höher als die Effektstärken in den anderen genannten Metaanalysen zu verschiedenen Kompetenzen. Edele et al. (2015) gehen jedoch davon aus, dass dieses Ergebnis nicht ohne weiteres generalisiert werden kann, da die Metaanalyse nur wenige Studien umfasst und überwiegend auf Fremdsprachenlernerinnen und -lerner sowie kleine, selektive Stichproben beschränkt ist. Neubauer und Hofer (2019) argumentieren, dass die Validität der Selbsteinschätzungen in dieser Metaanalyse deshalb höher sein könnte, weil Fremdsprachkompetenzen eingeschätzt wurden und Individuen beim Lernen einer Fremdsprache häufig Feedback bekommen. Edele et al. (2015) führten selbst ein Review zu den Korrelationen zwischen Selbsteinschätzungen von Sprachkompetenzen und getesteten Sprachkompetenzen durch. Sie bezogen Studien ein, die sowohl Selbsteinschätzungsmaße als auch Testleistungen oder Einschätzungen von Lehrkräften enthielten. Die meisten Stichproben dieser Studien waren Studierende, die eine Fremdsprache lernten. Weniger Studien untersuchten die Selbsteinschätzungen von Zweitsprachlernerinnen und -lernern, für die die Sprache im Aufenthaltsland die Zweitsprache darstellte. Darüber hinaus wurden wenige Studien einbezogen, in denen Erstsprachfähigkeiten betrachtet wurden. Insgesamt gab es wenige Studien mit heterogeneren oder jüngeren Stichproben. Über die verschiedenen Studien hinweg fanden Edele et al. (2015), dass die Korrelationskoeffizienten stark variierten. Als Begründung dafür sehen sie die Unterschiede in der Qualität der verwendeten Instrumente. Es wurden sowohl etablierte als auch fragwürdige Sprachkompetenztests verwendet als auch die Einschätzungen von Lehrkräften. Auch die Items zur Selbsteinschätzung unterschieden sich. Manche Studien verwendeten allgemeinere Selbsteinschätzungsitems, bei denen die Sprachkompetenz in verschiedenen Dimensionen (z.B. Verstehen, Sprechen, Lesen oder Schreiben) auf einer vier- oder fünfstufigen Antwortskala eingeschätzt werden sollte (z.B. Brantmeier, 2006; Finnie & Meng, 2005). In der Mehrheit der Studien wurden jedoch spezifischere Selbsteinschätzungsitems verwendet, die sich auf konkrete Kriterien beziehen, wie z.B. alltägliche Situationen (z.B. Brantmeier et al., 2012) oder auf konkrete Aufgaben, wie die Anzahl richtiger Antworten in einem Test (z.B. Lin et al., 2001). Für die wenigen Studien, die allgemeinere Selbsteinschätzungsitems verwendeten, waren die Ergebnisse gemischt. Während z.B. Brantmeier (2006) keinen signifikanten Zusammenhang zwischen Selbsteinschätzungen und Testergebnissen fand, lag die Korrelation zwischen Selbsteinschätzungen und Tests einer kanadischen Studie für im Ausland geborene Personen bei $r = .62$ (Finnie & Meng, 2005). In den Studien, in denen spezifischere Selbsteinschätzungen verwendet

wurden, wurde überwiegend von mittleren bis sehr hohen Korrelationen zwischen den verschiedenen Maßen berichtet. Besonders hoch ($r = .80$) waren die Korrelationen, wenn Fremdsprachlernerinnen und -lerner mit sehr kontextspezifischen langen Itemskalen befragt wurden (LeBlanc & Painchaud, 1985; vgl. Abschnitt 1.7.2).

Edele et al. (2015) analysierten zudem Daten des Migrationssamples der vierten Startkohorte des Nationalen Bildungspanels (NEPS). Sie fanden bei der Selbsteinschätzung der Deutschkompetenzen der Neuntklässlerinnen und Neuntklässler mit Migrationshintergrund einen Deckeneffekt. Der Median lag über verschiedene Herkunftsgruppen und die linguistischen Dimensionen Sprechen, Verstehen, Lesen und Schreiben hinweg mit einer Ausnahme bei der von fünf Antwortkategorien höchstmöglichen Antwortkategorie *sehr gut*. Die Mehrheit der Schülerinnen und Schüler wählten die Antwortkategorien *eher gut* und *sehr gut*. Die Variabilität der Selbsteinschätzungen war deshalb geringer als die der Testscores. Die Korrelationen zwischen Testscores zur deutschen Lesekompetenz und den Selbsteinschätzungen hatten kleine bis mittlere Effektstärken und lagen für das allgemein formulierte Item zum Lesen im Bereich von $r = .22$ bis $r = .31$. Weiterhin bearbeiteten Edele et al. (2015) die Fragestellung, ob die in der Migrationsforschung häufig genutzten Selbsteinschätzungen zu ähnlichen Ergebnissen führen, wie die Ergebnisse von Kompetenztests. Dazu korrelierten sie nicht nur die Selbsteinschätzungen mit den Testscores, sondern verglichen auch Ergebnisse von sonst identischen Regressionsanalysen, die sich nur im verwendeten Kompetenzmaß (Selbsteinschätzung oder Testscore) unterschieden. Zum Beispiel wurden in die Regressionsanalysen Determinanten von Sprachkompetenzen als unabhängige Variablen aufgenommen, während das jeweilige Kompetenzmaß die abhängige Variable darstellte. Zusammenfassend wurden in allen verglichenen Analysen in Abhängigkeit vom Kompetenzmaß unterschiedliche Effekte gefunden und aus dem Muster der Abweichungen schlussfolgerten die Autorinnen und der Autor, dass unter der Verwendung von Selbsteinschätzungen als Kompetenzmaß das Risiko, relevante Effekte zu übersehen, erhöht sei.

In weiteren Studien wurden ebenso die Selbsteinschätzungen von Sprachkompetenzen untersucht. Gollan et al. (2012) verglichen die Selbsteinschätzungen von 52 bilingualen Personen (Spanisch und Englisch) mit Interviewerratings und Scores von zwei Benennungstests und fanden Korrelationen zwischen $r = .28$ und $r = .50$ für Englisch und zwischen $r = .43$ und $r = .52$ für Spanisch. Darüber hinaus berechneten sie Scores für die Sprachdominanz, indem sie den Spanischscore vom Englischscore subtrahierten. Die Korrelationen zwischen der Sprachdominanz, die mit den Selbsteinschätzungen berechnet wurde und der Sprachdominanz, die mit den Interviewerratings oder Testscores berechnet wurde, lagen im Bereich von $r = .59$ bis $r = .62$. Die Korrelationen zur Sprachdominanz waren also etwas höher als die zu den Sprachkompetenzen. Trofimovich et al. (2016) untersuchten die Aussprache von 134 internationalen Studierenden in der Zweitsprache Englisch. Sie fanden keinen signifikanten Zusammenhang zwischen Selbsteinschätzungen ihres Akzents

und der Einschätzung ihres Akzents durch drei Personen mit englischer Muttersprache, die Audiodateien angehört hatten, in denen die Studierenden eine Bildergeschichte beschrieben ($r = .06$; $p = .50$). Zur Verständlichkeit fanden sie einen schwachen Zusammenhang zwischen Selbsteinschätzung und Fremdeinschätzung ($r = .18$; $p = .03$).

Insgesamt variieren die Ergebnisse von Studien zur Validität von Selbsteinschätzungen von Sprachkompetenzen. Mittlere bis große Effektstärken scheinen möglich zu sein, aber auch Nulleffekte können vorkommen. Die beschriebenen Studien wurden anhand verschiedener Stichproben durchgeführt. Dazu gehörten Migrantinnen und Migranten der ersten, zweiten und dritten Generation (Edele et al., 2015; Gollan et al., 2012) und internationale Studierende in verschiedenen Programmen (Trofimovich et al., 2016). Außerdem haben sich die Teilnehmenden im Alter und kulturellen Hintergrund unterschieden. Verschiedene Gruppen könnten Selbsteinschätzungen unterschiedlich beantworten. Die berichteten Ergebnisse können also nicht ohne weiteres auf andere Stichproben übertragen werden. Weiterhin unterscheiden sich die Studien in der betrachteten Kompetenz und Kompetenzdomäne. Darüber hinaus könnten auch methodische Aspekte zur Varianz zwischen den Studien beigetragen haben.

1.10 Zusammenfassung des theoretischen und empirischen Hintergrunds

Als nächstes fasse ich die Ausführungen zum theoretischen und empirischen Hintergrund der Arbeit zusammen, bevor ich im folgenden Kapitel 2 genauer auf die Fragestellung und die aus dem beschriebenen Hintergrund abgeleiteten Hypothesen eingehe.

Aus der beschriebenen Literatur ist bekannt, dass Selbsteinschätzungen diejenigen Teile des Wissens über eigene Fähigkeiten widerspiegeln, welche zum Zeitpunkt der Selbsteinschätzung im Working-Self-Concept verfügbar sind (s. Kapitel 1.1). Dabei wird das Selbstkonzept durch soziale Rückmeldungen und reflektierte Beurteilungen, Selbstwahrnehmung und soziale Vergleiche gespeist. Externale und internale Referenzrahmen, also Informationen über die Fähigkeiten anderer Personen oder eigene Fähigkeiten im ideationalen, temporalen oder dimensionalen Vergleich dienen zur Einordnung der selbstbezogenen Informationen. Interindividuell unterschiedliche Referenzrahmen, wie im Shifting-Standards-Modell, dem BFLPE und dem I/E-Modell beschrieben, beeinträchtigen die interindividuelle Vergleichbarkeit von Selbsteinschätzungen (s. Kapitel 1.2).

Darüber hinaus werden Selbsteinschätzungen durch verschiedene Motive beeinflusst. Das Self-Enhancement-Motiv beschreibt die Tendenz zu einem erhöht positiven Selbstbild, welches durch verschiedene Strategien aufrechterhalten wird. Das Self-Verification-Motiv beschreibt die Tendenz, das bestehende Selbstbild zu erhalten und zu bestätigen und das Self-Assessment-Motiv beschreibt die Tendenz, eigene Fähigkeiten möglichst genau einschätzen zu wollen. Das Self-

Enhancement-Motiv tritt insbesondere dann auf, wenn eine persönlich wichtige oder allgemein erstrebenswerte Domäne eingeschätzt wird und wenn es sich um eine uneindeutige Fähigkeit handelt, sodass Interpretationsspielraum gegeben ist. Es dominiert auch, wenn kognitive Ressourcen bei der Einschätzung knapp sind oder wenn ein nicht fest verankerter Bereich des Selbstbilds eingeschätzt wird. Sofern hingegen ein sicherer, fest verankerter Bereich des Selbstbilds eingeschätzt wird und ausreichend kognitive Ressourcen zur Verfügung stehen, um das Selbstbild abzurufen, dominiert das Motiv, das bestehende Selbstbild zu bestätigen. Zu einer möglichst genauen Selbsteinschätzung sind Personen dann motiviert, wenn die Fähigkeit eindeutig definiert ist und kein Interpretationsspielraum gegeben ist, wenn sie sich für ihre Einschätzung rechtfertigen müssen, wenn die genaue Selbsteinschätzung dazu dient, die eigene Fähigkeit zu verbessern oder wenn eine genaue Einschätzung aus einem anderen Grund wichtig ist (s. Kapitel 1.3).

Weiterhin werden Selbsteinschätzungen durch den kognitiven Entwicklungsstand beeinflusst. Insbesondere in der Kindheit sind Selbsteinschätzungen auch aufgrund von Limitationen der kognitiven Entwicklungsstufe unrealistisch positiv und werden mit voranschreitender kognitiver Entwicklung in der Kindheit und Jugend realistischer. Im mittleren Jugendalter sind Personen bereits dazu in der Lage, einzelne Eigenschaften in Konzepten höherer Ordnung zu integrieren und Abstraktionen zu bilden und können diese Abstraktionen miteinander vergleichen. Sie haben jedoch noch Schwierigkeiten, verschiedene Abstraktionen sinnvoll zu integrieren, sodass sich mögliche Widersprüche lösen. Diese Fähigkeit entwickelt sich erst im späten Jugendalter ab einem Alter von ca. 17 Jahren. Das heißt, erst in diesem Alter wird ein integriertes und kohärentes Selbstbild gefestigt (s. Kapitel 1.4).

Die Konstruktion des Selbstbilds unterscheidet sich zwischen individualistisch und kollektivistisch geprägten Kulturen. Personen mit individualistischer kultureller Prägung neigen zur Konstruktion eines independenten Selbstbilds. Der Fokus liegt auf Unabhängigkeit und persönlichen Zielen und die persönliche Identität muss kontinuierlich verifiziert und verteidigt werden, um ein kohärentes und stabiles Selbstbild zu erhalten. Dabei werden auch Self-Enhancement-Strategien angewendet. Personen mit kollektivistischer kultureller Prägung neigen zur Konstruktion eines interdependenten Selbstbilds. Der Fokus liegt auf dem Erreichen von Gruppenzielen und die Identität ist an soziale Rollen gebunden und muss weniger verteidigt oder hinterfragt werden. In kollektivistisch geprägten Kulturen kommt es nur unter bestimmten Umständen zum Self-Enhancement. Die deutsch- und englischsprachige Literatur konzentriert sich jedoch auf westliche und ostasiatische Kulturen. Inwiefern Kenntnisse zu Selbsteinschätzungen auch für Gruppen mit anderem kulturellem Hintergrund, wie z.B. Geflüchtete aus Syrien, Irak und Afghanistan, angenommen werden können, ist unklar (s. Kapitel 1.5).

Den Prozess von der Eigenschaft einer Person zu einer genauen Einschätzung dieser Eigenschaft beschreibt das RAM. Nach dem RAM müssen relevante Informationen über die Eigenschaft

der Person verfügbar sein, von der urteilenden Person wahrgenommen und schließlich korrekt interpretiert werden, um eine genaue Einschätzung vorzunehmen. Das Modell lässt sich auch auf Selbsteinschätzungen von Kompetenzen anwenden, wobei im Fall von Selbsteinschätzungen tendenziell mehr relevante Informationen verfügbar sind als bei Fremdeinschätzungen. Aus dem RAM lassen sich Moderatoren der Urteilsgenauigkeit ableiten und systematisieren. Einzelne Moderatoren können dem Judge, dem Target, dem Trait, der Information oder der Interaktion aus zwei oder mehr dieser übergeordneten Moderatoren zugeordnet werden. Auf Seiten des Judges wirken sich z.B. dessen Wissen, insbesondere dessen Expertise hinsichtlich des Traits, dessen Fähigkeiten, insbesondere die kognitiven Fähigkeiten und dessen Motivation zu einer genauen Einschätzung auf die Urteilsgenauigkeit aus. Auf Seiten des Targets ist es entscheidend für die Urteilsgenauigkeit, wie viele relevante Informationen die Person verfügbar macht und ob diese konsistent sind. Auf Seiten des Traits beeinflussen die Sichtbarkeit, die Wichtigkeit und die Eindeutigkeit die Urteilsgenauigkeit. Auf Seiten der Informationen spielen die Quantität und die Qualität eine entscheidende Rolle für die Genauigkeit des Urteils (s. Kapitel 1.6).

Die Beantwortung einer Selbsteinschätzungsfrage in einer Befragungssituation stellt die letzte Hürde für die Genauigkeit der Einschätzung dar. In der konkreten Situation muss das Item verstanden und die relevante Einschätzung aus dem Gedächtnis abgerufen werden. Liegt noch keine integrierte Einschätzung vor, müssen für die Einschätzung relevante Informationen abgerufen und integriert werden. Zuletzt muss eine zu der Einschätzung passende Antwortkategorie des Items gewählt werden. Wie gründlich Befragte bei der Beantwortung von Fragebogenitems vorgehen, wurde auf einem Kontinuum von Optimizing bis Strong Satisficing beschrieben. Idealerweise nehmen sie den Aufwand auf sich, die beschriebenen kognitiven Prozesse gründlich auszuführen. Entscheidend für die Gründlichkeit bei der Bearbeitung ist erstens die Motivation der Teilnehmenden, was z.B. von deren Erschöpfung abhängt. Zweitens ist die Aufgabenschwierigkeit entscheidend, also z.B. die Komplexität der Fragestellung oder situative Faktoren wie äußere Ablenkungen. Drittens sind Fähigkeiten der befragten Person entscheidend, insbesondere kognitive Fähigkeiten. Hinsichtlich der Itemgestaltung ist die Eindeutigkeit und Verständlichkeit der Frage und der Antwortkategorien wichtig, um reliable und valide Ergebnisse zu erzielen. Es kann vorteilhaft sein, eine Referenzgruppe vorzugeben und möglichst vertraute Fähigkeiten einschätzen zu lassen bzw. die Frage auf bekannte Situationen zu beziehen (s. Kapitel 1.7).

Die Erfassung der Genauigkeit von Selbsteinschätzungen kann wie eine Testvalidierung anhand verschiedener Kriterien erfolgen. Dabei ist es wichtig, dass die Kompetenzdomäne, die mit dem Kriterium gemessen wird, mit der Kompetenzdomäne übereinstimmt, nach der in der Selbsteinschätzung gefragt wird. Wird in der Selbsteinschätzung z.B. nach der allgemeinen deutschen Sprachkompetenz gefragt und die Genauigkeit dieser Einschätzung an einem Kompetenztest gemessen, der nur den rezeptiven Wortschatz im Deutschen erfasst, wird die Genauigkeit der

Selbsteinschätzung ggf. unterschätzt. Um die Urteilsgenauigkeit zu quantifizieren, können verschiedene Indikatoren berechnet werden, die sich auf folgende drei Komponenten der Genauigkeit beziehen: die Diskrimination, die allgemeine Verzerrung und die Variation. Die Diskrimination beschreibt, inwiefern die Selbsteinschätzung die konkrete Rangordnung innerhalb der untersuchten Gruppe wiedergibt. Dazu wird in der Regel die Korrelation zwischen Selbsteinschätzung und Kriterium berechnet. Die allgemeine Verzerrung beschreibt den allgemeinen Fehler, den Personen bei der Selbsteinschätzung machen, ob also eine allgemeine Über- oder Unterschätzung vorliegt, welche von der Korrelation zwischen Selbsteinschätzung und Kriterium statistisch unabhängig ist. Die letzte beschriebene Komponente stellt die Variation der Werte innerhalb der Gruppe dar, also inwiefern die Variation der Kompetenz durch die Selbsteinschätzung richtig erfasst wird (s. Kapitel 1.8).

Metaanalytisch wurden Korrelationen mittlerer Effektstärke zwischen Selbsteinschätzungen verschiedener Kompetenzen und Kriterien zur Validierung gefunden. Die Korrelationen zwischen Selbsteinschätzungen von Sprachkompetenzen und verschiedenen Kriterien variierten sehr stark. Während teilweise kein Zusammenhang gefunden wurde, lagen viele Korrelationen im mittleren bis hohen Bereich. Ein großer Teil der untersuchten Stichproben waren Fremdsprachenlernende. Im Migrationssample der Startkohorte 4 des NEPS fand sich bei Selbsteinschätzungen der deutschen Sprachkompetenz ein Deckeneffekt, sodass die Variabilität der Selbsteinschätzungen die Variabilität der Ergebnisse von Kompetenztests nicht angemessen wiedergeben konnte. Auch die Korrelationen zwischen den Selbsteinschätzungen und Testergebnissen zeigten nur kleine bis mittlere Effektstärken. Ergebnisse von Analysen zu verschiedenen Themen unterschieden sich in dieser Studie in Abhängigkeit vom verwendeten Kompetenzmaß. Insgesamt variieren die Befunde zur Genauigkeit von Selbsteinschätzungen und unterscheiden sich u.a. zwischen verschiedenen untersuchten Gruppen und Kompetenzen bzw. Kompetenzdomänen und auch abhängig von den verwendeten Instrumenten. Zur Genauigkeit von Selbsteinschätzungen der Deutschkompetenzen jugendlicher Flüchtlinge liegen meines Wissens noch keine Ergebnisse vor (s. Kapitel 1.9).

2 Fragestellung und Hypothesen

Ziel meiner Arbeit ist die Untersuchung der Selbsteinschätzungen der sprachlichen Kompetenzen im Deutschen jugendlicher Flüchtlinge. Damit betrachte ich eine in mehrfacher Hinsicht besondere Gruppe, bei der zwei Aspekte besonders relevant für die Selbsteinschätzungen sind. Erstens befanden sich die Teilnehmenden im mittleren Jugendalter, sodass das Selbstbild in der Regel noch durch nicht integrierte Widersprüche geprägt und weniger kohärent und gefestigt war als im späten Jugend- oder Erwachsenenalter (s. Kapitel 2.4). Zweitens handelte es sich um überwiegend aus Syrien, aber auch aus weiteren Ländern stammende geflüchtete Personen, die zum Zeitpunkt der Befragung mit Familienangehörigen in Deutschland lebten. Die Personen hatten einen anderen kulturellen Hintergrund als die meisten Teilnehmenden oben beschriebener empirischer Untersuchungen, lebten jedoch in Deutschland und waren bereits mehr oder weniger vertraut mit der deutschen Kultur und den deutschen Wertvorstellungen (s. Kapitel 1.5).

Im Kapitel zum theoretischen und empirischen Hintergrund dieser Arbeit ist beschrieben, welche Quellen und Referenzen für selbstbezogene Informationen herangezogen werden, um zu einer Selbsteinschätzung zu gelangen, welche Motive die Selbsteinschätzungen beeinflussen können, welche Rolle die altersbedingte kognitive Entwicklung für die Entwicklung von Selbsteinschätzungen spielt und wie diese durch kulturelle Prägungen beeinflusst werden. Das RAM beschreibt verschiedene Stufen des Prozesses, der zu einer genauen Selbsteinschätzung führen kann und es können Moderatoren der Urteilsgenauigkeit daraus abgeleitet werden. Weiterhin wurde auf die Operationalisierung von Selbsteinschätzungsitems und die kognitiven Prozesse bei der Beantwortung von Selbsteinschätzungsfragen sowie auf die Gestaltung von Selbsteinschätzungsfragen eingegangen. Nach der Erläuterung methodischer Aspekte der Erfassung der Genauigkeit von Selbsteinschätzungen wurden empirische Befunde zur Genauigkeit von Selbsteinschätzungen zusammengefasst.

Es hat sich gezeigt, dass die Befunde zur Genauigkeit von Selbsteinschätzungen heterogen sind und häufig an speziellen Stichproben und speziellen Umständen (z.B. im Rahmen eines Sprachkurses) geprüft wurden, sodass diese nicht ebenso für die Genauigkeit der Selbsteinschätzungen von jugendlichen Flüchtlingen in Deutschland angenommen werden können, sondern die Gültigkeit bisheriger Befunde für diese Gruppe noch untersucht werden muss. Zur Untersuchung der Selbsteinschätzungen dieser Gruppe stelle ich zunächst die Frage danach, wie die jugendlichen Geflüchteten ihre Deutschkompetenz selbst einschätzen und wie die Einschätzungen mit objektiv gemessenen Kompetenzen zusammenhängen und übereinstimmen. Dies beinhaltet die Frage nach der Genauigkeit und Validität der Selbsteinschätzungen der Deutschkompetenzen der jugendlichen Flüchtlinge.

Eine weitere Fragestellung bezieht sich darauf, durch welche Faktoren die Selbsteinschätzungen der jugendlichen Flüchtlinge beeinflusst werden. In der oben beschriebenen Literatur wurden in der Regel einzelne Quellen selbstbezogener Informationen betrachtet, oder einzelne Referenzrahmen, Motive und Verzerrungen empirisch untersucht. Im Rahmen des RAM wurden Moderatoren von Einschätzungen auf theoretischer Ebene systematisiert. Dabei bezieht sich das RAM ursprünglich auf Fremdeinschätzungen von Persönlichkeitseigenschaften und die daraus abgeleitete Forschung zu Moderatoren der Urteilsgenauigkeit hat denselben Fokus (z.B. Colman, 2021). Entsprechend gibt es bisher meines Wissens keine die verschiedenen Moderatoren von Selbsteinschätzungen integrierenden empirischen Studien und keine Studien zu den Moderatoren von Selbsteinschätzungen von Deutschkompetenzen jugendlicher Flüchtlinge in Deutschland im Speziellen. Um Aussagen über die Gültigkeit und Übertragbarkeit des theoretischen und empirischen Hintergrunds zu generieren und das bestehende Wissen zu replizieren, zu ergänzen und zu erweitern, möchte ich in dieser Arbeit verschiedene Einflussfaktoren der Selbsteinschätzungen mit einem integrierenden Ansatz empirisch nachweisen und so die Zusammensetzung verschiedener Einflussfaktoren auf die Selbsteinschätzungen der Sprachkompetenzen jugendlicher Flüchtlinge empirisch untersuchen.

Die letzte übergeordnete Fragestellung bezieht sich auf die Vor- und Nachteile von verschiedenen Selbsteinschätzungsitems. Methodische Aspekte der Genauigkeit von Selbsteinschätzungen von Kompetenzen wurden bisher vor allem metaanalytisch verglichen, also indem die Validität von Selbsteinschätzungen zwischen Studien verglichen wurde, die verschiedene Arten von Items einsetzten. Dabei wurden Selbsteinschätzungen z.B. danach kategorisiert, ob sie einen Vergleich instruieren oder eine Referenzgruppe vorgeben. Auf diese Weise können jedoch die Unterschiede in Studiensettings, Teilnehmergruppen und objektiven Kriteriumsmaßen kaum kontrolliert werden. Ein direkter Vergleich von Selbsteinschätzungsitems innerhalb einer Studie anhand derselben Stichprobe hat den Vorteil, dass bestimmte Aspekte von Items systematisch variiert werden können, während das Studiensetting, die Stichprobe und das objektive Kompetenzmaß gleichbleiben. Somit kann ein besser kontrollierter Vergleich der Items erfolgen. In dieser Arbeit vergleiche ich vier verschiedene Selbsteinschätzungsskalen miteinander, die sich aufgrund empirischer Erfahrungen und theoretischer Überlegungen in bestimmten Aspekten voneinander unterscheiden. Ich bearbeite dazu die Fragen, welche Items die Deutschkompetenzen der Teilnehmenden am genauesten erfassen und welche Vor- und Nachteile bestimmte Items bieten.

Mit der Beantwortung der beschriebenen Fragestellungen trage ich neben dem theoretischen Interesse auch zu der einleitend beschriebenen praktischen Problematik bei der Entscheidung zum Einsatz von Selbsteinschätzungen oder objektiven Kompetenzmaßen bei. Die praktische Frage bezieht sich darauf, inwiefern die Deutschkompetenzen der jugendlichen Flüchtlinge mit Selbsteinschätzungen erfasst werden können, sodass diese eine Alternative für Kompetenztestungen

darstellen. Einen Mehrwert für (potenzielle) Anwenderinnen und Anwender von Selbsteinschätzungsitems soll darüber hinaus die Identifikation der Variablen bringen, welche mit Selbsteinschätzungen konfundiert sein könnten, sodass dies bei der Verwendung von Selbsteinschätzungen als Kompetenzmaß bzw. der Interpretation der Ergebnisse berücksichtigt werden kann. Zuletzt bietet der Vergleich verschiedener Selbsteinschätzungsitems Informationen für die Itemauswahl zukünftiger Anwendungen von Selbsteinschätzungen, wodurch möglicherweise die Genauigkeit von Selbsteinschätzungen als Kompetenzmaß verbessert werden kann.

Zur Beantwortung der Fragestellungen verwende ich in Kapitel 3 näher beschriebene Daten der Studie *Refugees in the German Educational System* (ReGES), welche u.a. Befragungsdaten und Testdaten von überwiegend aus Syrien stammenden jugendlichen Geflüchteten in Deutschland zu zwei Messzeitpunkten umfassen. Die Befragungsdaten enthalten zum einen Einschätzungen auf vier verschiedenen Arten von Selbsteinschätzungsskalen zu den Deutschkompetenzen. Weiterhin enthalten die Befragungsdaten verschiedene Informationen über die Jugendlichen, die mögliche Moderatoren der Selbsteinschätzungen der Deutschkompetenzen darstellen. Getestet wurde das Hörverstehen der deutschen Sprache in den Bereichen Wortschatz und Grammatik, weshalb in dieser Arbeit hauptsächlich die Selbsteinschätzungen zum Verstehen der deutschen Sprache betrachtet werden, ggf. werden auch Selbsteinschätzungen zum Sprechen der deutschen Sprache in Analysen miteinbezogen (s. Abschnitt 3.3.1.1). Außerdem wurden kognitive Grundfähigkeiten der Jugendlichen getestet.

In den drei folgenden Unterkapiteln leite ich anhand des in Kapitel 1 beschriebenen theoretischen und empirischen Hintergrunds Hypothesen zu den übergeordneten Fragestellungen her. Zuerst gehe ich auf Hypothesen zur Genauigkeit und Validität der Selbsteinschätzungen ein, dann auf Hypothesen zu den Faktoren, die die Selbsteinschätzungen der Deutschkompetenzen beeinflussen, und zuletzt auf Hypothesen zum Vergleich der verschiedenen Selbsteinschätzungsitems.

2.1 Hypothesen zur Genauigkeit der Selbsteinschätzungen

Zur Beantwortung der Frage nach der Genauigkeit der Selbsteinschätzungen der jugendlichen Flüchtlinge ziehe ich die drei in Kapitel 1.8 genannten Indikatoren heran: die Diskrimination, die allgemeine Verzerrung und die Variation. Für jeden Indikator leite ich eine Hypothese aus dem theoretischen und empirischen Hintergrund ab und prüfe diese.

2.1.1 Diskrimination

Die Diskrimination bezieht sich auf die Korrelation zwischen Selbsteinschätzung und tatsächlicher Kompetenz. Inwiefern die Selbsteinschätzungen die korrekte Rangordnung der

Deutschkompetenz innerhalb einer untersuchten Gruppe wiedergeben können, wird aus theoretischer Sicht durch verschiedene Aspekte erschwert und durch andere begünstigt. Internale und externale Referenzrahmen unterscheiden sich zwischen Teilnehmenden, sodass z.B. der BFLPE und internale Vergleiche wie im I/E-Modell beschrieben die interindividuelle Vergleichbarkeit der Selbsteinschätzungen beeinträchtigen können (vgl. Kapitel 1.2).

Weiterhin ist im mittleren Jugendalter die Entwicklung kognitiver Strukturen, die für die Konstruktion eines realistischen und kohärenten Selbstbilds relevant sind, bereits fortgeschritten. Jedoch ist die Fähigkeit zur Integration von Abstraktionen höherer Ordnung noch eingeschränkt, sodass Widersprüche nicht gelöst werden können (vgl. Kapitel 1.4). Möglicherweise ist das Selbstbild also noch nicht so gefestigt, dass Informationen zur Selbsteinschätzung einfach aus dem Selbstkonzept abgerufen werden können und Urteile zu einem größeren Anteil ad hoc gebildet werden müssen, was die Selbsteinschätzung erschwert und wodurch diese maßgeblich durch die aktuelle Zugänglichkeit zu relevanten Gedächtnisinhalten beeinflusst wird (vgl. Abschnitt 1.7.1), woraus tendenziell ungenauere Selbsteinschätzungen resultieren.

Darüber hinaus ist grundsätzlich davon auszugehen, dass die Neigung zum Self-Enhancement interindividuell unterschiedlich stark ausgeprägt ist. Dabei kann u.a. die individuell unterschiedliche Wichtigkeit der Deutschkompetenz die individuelle Neigung zum Self-Enhancement beeinflussen (vgl. Abschnitt 1.3.1). Interindividuelle Unterschiede in der Neigung zum Self-Enhancement beeinträchtigen die Diskriminationsfähigkeit der Selbsteinschätzungen, denn nur, wenn sich alle Personen gleichermaßen über- oder unterschätzen, bleibt die Diskriminationsfähigkeit des Selbsteinschätzungsmaßes erhalten.

Weiterhin kann die Gründlichkeit, mit der Personen Selbsteinschätzungen beantworten durch die kognitive Belastung der Befragungssituation beeinflusst sein (vgl. Abschnitt 1.7.1). In der ReGES-Studie hatten die Teilnehmenden bereits einen relativ langen Fragebogen beantwortet, bis sie zu den Selbsteinschätzungsfragen zur Sprachkompetenz gelangten. Die kognitive Belastung war demnach hoch, weshalb von einer weniger gründlichen Beantwortung auszugehen ist.

Außerdem ist die Eindeutigkeit des einzuschätzenden Traits relevant für die Genauigkeit der Selbsteinschätzungen (vgl. Abschnitt 1.3.1). Denn wenn Personen unterschiedliche Aspekte des Traits bewerten, ist die Vergleichbarkeit beeinträchtigt. Hinsichtlich der Eindeutigkeit bietet die Sprachkompetenz insofern Interpretationsspielraum, dass sie eine breite Kompetenz ist, die viele unterschiedliche Dimensionen und Kompetenzbereiche umfasst. Es kann u.a. zwischen Fähigkeiten im Verstehen, Sprechen, Lesen und Schreiben sowie zwischen grammatischer Kompetenz, Aussprache, Wortschatz, etc. unterschieden werden. In der Formulierung von Selbsteinschätzungsitems können solche Kompetenzbereiche genauer definiert sein. Da bei den verwendeten Items jede Dimension (Verstehen, Sprechen, Lesen, Schreiben) einzeln bewertet wird, der Kompetenzbereich jedoch nicht weiter definiert ist, ist zumindest hinsichtlich der Kompetenzbereiche von

einem interindividuell unterschiedlichen Fokus auszugehen, was die Diskrimination der Selbsteinschätzungsmaße in entsprechendem Ausmaß beeinträchtigen kann.

Auf Seiten des Traits ist das Vorkommen relevanten Verhaltens und somit die Verfügbarkeit von Informationen relevant für die Genauigkeit der Selbsteinschätzungen (vgl. Abschnitte 1.6.1.1 und 1.6.2.3). Vor dem Hintergrund des Zusammenhangs zwischen Sprachkontakt und Sprachkompetenz (z.B. Chiswick & Miller, 2001) sollten in der Regel relevante Informationen verfügbar sein, falls mindestens ein geringes Kompetenzniveau vorliegt. In jeder Sprachkontaktsituation, in der Personen eine Sprache erwerben, sollten relevante Hinweise auf deren Sprachkompetenz verfügbar sein. Es sollten z.B. Hinweise verfügbar sein, wie gut sie die gesprochene oder geschriebene Sprache verstehen oder wie leicht es ihnen fällt, sich schriftlich oder mündlich auszudrücken. Falls es keinen Sprachkontakt gab, ist die Einschätzung deshalb einfach, weil dann auch keine Kompetenz in der Sprache vorhanden sein kann, mit wenigen Ausnahmen, wenn jemand z.B. eine verwandte Sprache beherrscht und dadurch in der Lage ist, die deutsche Sprache teilweise zu verstehen, obwohl er oder sie zu dieser bisher keinen Kontakt hatte. Dieser Zusammenhang zwischen relevanten Informationen (Sprachkontakt) und Ausprägung der zu beurteilenden Eigenschaft (sprachliche Kompetenz), ist eine Besonderheit von Kompetenzen, insbesondere Sprachkompetenzen, die für Persönlichkeitseigenschaften nicht in diesem Ausmaß gilt. Es ist also zu vermuten, dass in der Regel ausreichend Informationen zur Verfügung stehen, um die eigene Deutschkompetenz genau einschätzen zu können.

In empirischen Studien wurden Zusammenhänge zwischen Selbsteinschätzungen und objektiven Kompetenzmaßen sprachlicher Kompetenzen im Deutschen in unterschiedlicher Höhe gefunden. Die Höhe der Zusammenhänge hing von den verwendeten Instrumenten und untersuchten Stichproben ab. In einem großen Teil der Studien wurden spezifische Selbsteinschätzungsmaße verwendet und für diese überwiegend mittlere bis hohe Korrelationen gefunden. Zu allgemeineren Selbsteinschätzungsmaßen gibt es wenige Studien und die Befunde sind gemischt. In einer großen Stichprobe Jugendlicher mit Migrationshintergrund wurden für allgemeine Selbsteinschätzungsmaße Korrelationen mittlerer Effektstärke mit objektiven Deutschkompetenzmaßen gefunden (s. Kapitel 1.9).

Aufgrund der zahlreichen Aspekte, die eine hohe Diskrimination von Selbsteinschätzungen beeinträchtigen können, wie den Referenzrahmeneffekten, Einschränkungen aufgrund der jugendlichen kognitiven Entwicklungsstufe und dem vermutlich noch nicht gefestigten Selbstkonzept, der möglicherweise unterschiedlich starken Neigung zum Self-Enhancement, der kognitiven Belastung der Teilnehmenden in der Befragungssituation und der Uneindeutigkeit hinsichtlich des zu beurteilenden Kompetenzbereichs ist kein hoher Zusammenhang zwischen Selbsteinschätzung und objektivem Kompetenzmaß zu erwarten. Die hohe Verfügbarkeit relevanter Informationen über die eigene Deutschkompetenz begünstigt jedoch eine genaue Einschätzung, sodass

zumindest ein kleiner Zusammenhang zwischen Selbsteinschätzungen und objektivem Kompetenzmaß realistisch erscheint. Für die jugendlichen Flüchtlinge wird deshalb ein positiver Zusammenhang geringer Stärke zwischen Selbsteinschätzungen und objektiven Kompetenzmaßen erwartet. Entsprechend lautet die Hypothese:

Hypothese 1a: Es besteht ein positiver Zusammenhang geringer Stärke zwischen den Selbsteinschätzungen der Deutschkompetenzen und den objektiven Kompetenzmaßen.

2.1.2 Allgemeine Verzerrung

Zur Beurteilung der allgemeinen Über- oder Unterschätzung ist zu definieren, wie die Skalenpunkte der Selbsteinschätzung zu interpretieren sind und wann eine allgemeine Über- oder Unterschätzung vorliegt. Wendet man das Shifting-Standards-Modell auf die Selbsteinschätzungen der Deutschkompetenzen von jugendlichen Flüchtlingen an, ist davon auszugehen, dass die jugendlichen Flüchtlinge bei der Einschätzung ihrer Deutschkompetenzen auf subjektiven Ratingskalen von dem Stereotyp ausgehen, dass kürzlich nach Deutschland zugewanderte Personen schlechtere Deutschkompetenzen haben als Personen, die seit längerer Zeit in Deutschland leben oder deren Muttersprache Deutsch ist. Demnach ist anzunehmen, dass sie subjektive Ratingskalen so interpretieren, dass sie die Verteilung der Deutschkompetenz der Geflüchteten in Deutschland abbildet. Berücksichtigt man auch den BFLPE bzw. welche Informationen über die Verteilung der Deutschkompetenz den Jugendlichen überhaupt zur Verfügung stehen, argumentiere ich, dass jugendliche Flüchtlinge insbesondere die Deutschkompetenzen der anderen jugendlichen Flüchtlinge in ihrer Schulklasse bzw. in ihrem Umfeld als Referenzrahmen auf subjektiven Ratingskalen annehmen und die Antwortskala so interpretierten, dass sie dieses Spektrum abbildet. Welche Skalenpunkte bei welcher objektiv gemessenen Kompetenz als Über- oder Unterschätzung gewertet werden, wird in den Abschnitten 3.3.2.3 und 4.4.2 definiert und erläutert.

Entscheidend dafür, ob eine allgemeine Verzerrung vorliegt, sind die allgemeine Neigung zum Self-Enhancement der Teilnehmenden und allgemeine kognitive Aspekte, die eine Verzerrung in eine bestimmte Richtung begünstigen. Die Neigung zum Self-Enhancement wird durch verschiedene Aspekte beeinflusst. Dazu gehören die Wichtigkeit, Eindeutigkeit und Überprüfbarkeit der einzuschätzenden Fähigkeit sowie die kognitive Belastung, die Stabilität des Selbstbilds, die kulturelle Prägung der Teilnehmenden und die Wichtigkeit einer genauen Einschätzung. Darüber hinaus kann der Dunning-Kruger-Effekt zu einer positiven Verzerrung der Selbsteinschätzung beitragen (vgl. Kapitel 1). Welche Ausprägungen der genannten Aspekte im vorliegenden Fall anzunehmen sind und wie sich diese ausgewirkt haben könnten, wird im Folgenden diskutiert und abschließend zu einer Hypothese zur allgemeinen Verzerrung der Selbsteinschätzungen der Deutschkompetenzen der jugendlichen Flüchtlinge zusammengefasst.

Personen sind insbesondere dann zum Self-Enhancement motiviert, wenn sie eine subjektiv zentrale und wichtige Kompetenz einschätzen (s. Kapitel 1.3.1). Deshalb muss diskutiert werden, inwiefern die Deutschkompetenz für jugendliche Flüchtlinge eine zentrale und wichtige Kompetenz darstellt. Es gibt einige Hinweise aus Ergebnissen der ReGES-Studie, die implizieren, dass die deutsche Sprachkompetenz eine wichtige Rolle für jugendliche Flüchtlinge spielt. Zum einen gaben 96% der Jugendlichen an, dass sie eine von verschiedenen Möglichkeiten im Alltag genutzt haben, um Deutsch zu lernen, wie z.B. die Nutzung des Internets auf Deutsch oder Gespräche mit Deutschen, was für ein gewisses Maß an Motivation zum Lernen der deutschen Sprache spricht. Wöchentlich bis täglich Zeit mit Deutschen zu verbringen gaben 89% der jugendlichen Flüchtlinge an, was dafürspricht, dass die deutsche Sprache eine wichtige Rolle in ihrem Alltag einnahm, vorausgesetzt sie kommunizierten auf Deutsch. Außerdem gaben 83% der befragten Jugendlichen an, für immer in Deutschland bleiben zu wollen, was die subjektive Relevanz der deutschen Sprache ebenfalls gestärkt haben sollte (Will et al., 2018). Es ist also anzunehmen, dass die Deutschkompetenz für die meisten jugendlichen Flüchtlinge eine erstrebenswerte, persönlich wichtige Kompetenz darstellt, was es aus dieser Hinsicht wahrscheinlich macht, dass sie bei der Einschätzung ihrer Deutschkompetenz zum Self-Enhancement motiviert sind.

Die Eindeutigkeit der einzuschätzenden Kompetenz wurde bereits in Abschnitt 2.1.1 zur Diskrimination diskutiert, mit dem Ergebnis, dass in den verwendeten Items zumindest die Sprachdimension (Verstehen, Sprechen, Lesen, Schreiben) vorgegeben ist, hinsichtlich des Kompetenzbereichs jedoch Interpretationsspielraum besteht. Somit sind eigennützige Interpretationen im Sinne des Self-Enhancements möglich.

Die Motivation zum Self-Enhancement ist gemindert, wenn Teilnehmende eine Überprüfung ihrer Selbsteinschätzung erwarten (vgl. Abschnitte 1.3.1 und 1.3.3). Sprachkompetenzen sind in Interaktionen grundsätzlich leicht überprüfbar, da zumindest für Personen mit hoher Kompetenz in der Sprache Fehler leicht zu identifizieren sind. In der ReGES-Studie ist jedoch nicht davon auszugehen, dass die Teilnehmenden eine Überprüfung ihrer Fähigkeiten erwarteten. Die Befragung erfolgte anonym und wurde in der Regel selbstständig an einem Tablet ausgefüllt. Darüber hinaus konnten Befragte nur in Ausnahmefällen bereits zum Zeitpunkt der Befragung wissen, dass die angekündigte Aufgabenbearbeitung Tests zu Deutschkompetenzen enthielt. Die wenigsten erwarteten also, dass ihre Angaben anhand objektiver Kompetenzdaten überprüft würden. Und tatsächlich hatte die Genauigkeit der Selbsteinschätzungen der Deutschkompetenz keinerlei Konsequenzen für die Teilnehmenden. Es ist also auch aus dieser Hinsicht keine Minderung des Self-Enhancement-Motivs zu erwarten.

Wie ebenfalls bereits im vorangegangenen Abschnitt zur Diskrimination argumentiert, könnte die Gründlichkeit bei der Beantwortung der Fragen durch die kognitive Belastung wegen des langen vorangegangenen Fragebogens beeinträchtigt gewesen sein und eine heuristische

Beantwortung begünstigt haben. Unter kognitiver Belastung dominiert in der Regel das Self-Enhancement-Motiv (vgl. Abschnitte 1.3.1.4 und 1.3.2).

Dass das Selbstbild von Jugendlichen altersbedingt tendenziell noch nicht gefestigt ist (vgl. Kapitel 1.4), begünstigt ebenfalls eine Dominanz des Self-Enhancement-Motivs insbesondere über das Self-Verification-Motiv, sodass zu erwarten ist, dass auch Personen mit eher negativem Selbstbild eine positive Ausprägung auf der Selbsteinschätzungsskala wählen (vgl. Abschnitt 1.3.2).

Hinsichtlich der kulturellen Herkunft sind jugendliche Geflüchtete vermutlich weniger individualistisch geprägt, als es bei Personen mit westlicher kultureller Prägung der Fall ist. Andererseits haben sie, wenn sie bereits seit einiger Zeit in Deutschland leben, auch mit der individualistischeren deutschen Kultur Erfahrungen gesammelt und die Werte wahrscheinlich in unterschiedlichem Ausmaß kennengelernt und für sich übernommen. Entsprechend ist auf Gruppenebene keine rein interdependente Konstruktion des Selbstbilds mit einem Hang zur Bescheidenheit zu erwarten, sondern auch eine independente Konstruktion des Selbstbilds mit einem Hang zum Self-Enhancement möglich (vgl. Kapitel 1.5).

In dem Ausmaß, in dem der Dunning-Kruger-Effekt auftritt, ist im Durchschnitt eine Überschätzung der eigenen Fähigkeiten zu erwarten. Das heißt, in den Kompetenzbereichen, in denen die einzuschätzende Kompetenz selbst relevant ist, um Fehler zu erkennen, neigen Personen zu einer Überschätzung ihrer Kompetenz, sofern die Kompetenz nicht ausreichend vorhanden ist. Wie in Abschnitt 1.6.2.1 erläutert, ist dies bei der Sprachkompetenz nicht in allen, aber in manchen Kompetenzbereichen der Fall. Insgesamt kann die Einschätzung der Sprachkompetenz durch den Dunning-Kruger-Effekt demzufolge positiv verzerrt werden.

Die diskutierten Aspekte sprechen alle für eine mehr oder weniger stark ausgeprägte Motivation der Teilnehmenden zum Self-Enhancement bzw. für eine positive Verzerrung der Selbsteinschätzungen. Es wird deshalb folgende Hypothese aufgestellt:

Hypothese 1b: Die jugendlichen Flüchtlinge überschätzen ihre Deutschkompetenz im Durchschnitt.

2.1.3 Variation

Es gibt verschiedene Gründe anzunehmen, dass die Variation von Selbsteinschätzungen geringer ausfällt als die Variation der tatsächlich zugrundeliegenden Kompetenz. Zunächst ist davon auszugehen, dass die Variation der Deutschkompetenzen von jugendlichen Flüchtlingen groß ist und sich im Bereich von keiner Kompetenz bis zum Muttersprachniveau bewegt. Diese Spannbreite mit einer fünfstufigen Skala angemessen abzubilden, erscheint fast unmöglich, da sehr große Bereiche zu Kategorien zusammengefasst werden müssen. Dazu kommt, dass die drei unteren Kategorien zumindest in der Migrationsstichprobe der Startkohorte 4 des NEPS kaum

gewählt wurden (Edele et al., 2015; s. Kapitel 1.9). Unter der Annahme, dass jugendliche Flüchtlinge zur Überschätzung ihrer Kompetenz neigen, stößt die Skala im oberen Bereich an ihre Grenze. Wählen die meisten Teilnehmenden eine der beiden oberen Kategorien, haben Personen mit weit überdurchschnittlicher Kompetenz nicht die Möglichkeit, sich von anderen Teilnehmenden durch eine angemessene Antwortalternative abzuheben. In dem Fall besteht ein Deckeneffekt, der die Variation der Selbsteinschätzungen einschränkt. Die gewählten Antwortalternativen sind im oberen Bereich der Skala zusammengestaucht. Es wird also angenommen, dass die Variation durch den Hang zur Überschätzung und die bevorzugte Wahl der oberen Antwortkategorien bei gleichzeitiger Beschränkung der Skala nach oben geringer ausfällt als die Variation der tatsächlichen Kompetenz. Die entsprechende Hypothese lautet:

Hypothese 1c: *Die Variation der Selbsteinschätzungen ist geringer als die Variation der objektiv gemessenen Kompetenzen.*

2.2 Hypothesen zu Einflussfaktoren der Selbsteinschätzungen

Zur Identifikation von Faktoren, die Selbsteinschätzungen inkrementell zur einzuschätzenden objektiven Kompetenz beeinflussen, werden im Folgenden Hypothesen aus dem theoretischen und empirischen Hintergrund abgeleitet. Die Hypothesen beziehen sich immer auf den direkten Effekt des jeweiligen Faktors auf die Selbsteinschätzung unter Kontrolle des Einflusses der objektiven Kompetenz auf die Selbsteinschätzung. Es werden auch Faktoren berücksichtigt, für die zunächst ein Effekt auf die Genauigkeit der Selbsteinschätzung vorhergesagt wird, welche in diesem Fall als der Betrag der Abweichung zwischen Selbsteinschätzung und objektiver Kompetenz, unabhängig von der Richtung der Abweichung, ob also eine Über- oder Unterschätzung vorliegt, definiert wird. Beeinflusst ein Faktor die Selbsteinschätzungen so, dass Über- und Unterschätzung in der Stichprobe gleichermaßen reduziert werden, würde nur ein Effekt auf die Genauigkeit der Selbsteinschätzungen gefunden, nicht jedoch ein Effekt in eine bestimmte Richtung auf die Selbsteinschätzungen an sich. Unter der in Abschnitt 2.1.2 erläuterten Annahme einer allgemeinen Überschätzung der Deutschkompetenzen wird davon ausgegangen, dass Faktoren, die zu genaueren Selbsteinschätzungen beitragen, einen negativen Effekt auf die Selbsteinschätzungen haben, weil die Reduktion der Überschätzung ein größeres Ausmaß hat als die Reduktion der Unterschätzung, in Summe die Selbsteinschätzung also nicht nur genauer, sondern auch in negativer Richtung beeinflusst wird. Aus diesem Grund wird auch für diese Faktoren ein Einfluss auf die Selbsteinschätzungen in eine bestimmte Richtung vorhergesagt, worauf in der Diskussion in Kapitel 5.2 eingegangen wird.

Als Einflussfaktoren werden die Leistung in Mathematik, die persönliche Wichtigkeit der Deutschkompetenz, Fähigkeiten zum schlussfolgernden Denken, die Teilnahme an Deutschunterricht und die Teilnahme an einem Deutschtest vorhergesagt. Zu weiteren Einflussfaktoren, wie z.B. der Sprachkompetenz des direkten Umfelds, die sich entsprechend dem BFLPE auf die Selbsteinschätzungen ausgewirkt haben könnte, werden keine Hypothesen aufgestellt, da eine entsprechende Überprüfung anhand der verwendeten Daten nicht möglich ist.

Nach dem I/E-Modell hat die mathematische Fähigkeit aufgrund des internalen dimensionalen Vergleichs einen direkten negativen Effekt auf die Selbsteinschätzung der Deutschkompetenz. Da sich der Effekt über verschiedene Kulturen als robust erwiesen hat, wird er auch in diesem Fall angenommen (vgl. Abschnitt 1.2.5). Es wird folgende Hypothese aufgestellt:

Hypothese 2a: Die Leistung in Mathematik hat einen negativen Effekt auf die Selbsteinschätzung der Deutschkompetenz.

Des Weiteren wird eine Auswirkung der persönlichen Wichtigkeit auf die Selbsteinschätzung der Deutschkompetenz vorhergesagt. In Abschnitt 2.1.2 zur allgemeinen Verzerrung wurde argumentiert, dass die jugendlichen Flüchtlinge insgesamt zum Self-Enhancement neigen, weil u.a. die Deutschkompetenz eine für die Gruppe wichtige Kompetenz darstellt und dies die Tendenz zum Self-Enhancement stärkt (vgl. Abschnitt 1.3.1). Ergänzend zu dieser vermuteten allgemeinen Tendenz, dass Deutschkompetenzen für jugendliche Flüchtlinge eine erstrebenswerte, wichtige Kompetenz darstellen, ist auch von individuellen Unterschieden auszugehen. Sicherlich ist es einigen jugendlichen Flüchtlingen wichtiger, die deutsche Sprache zu beherrschen, als anderen. Dies sollte sich insbesondere darin widerspiegeln, welches Engagement sie beim Erlernen der deutschen Sprache zeigen. Entsprechend dem in Abschnitt 1.3.1 beschriebenen Self-Centrality-Breeds-Self-Enhancement-Prinzip wird angenommen, dass sich die persönliche Wichtigkeit der Deutschkompetenz auch individuell auf die Neigung zum Self-Enhancement auswirkt. Jugendliche, für die ihre Deutschkompetenz eine wichtigere Rolle spielt und die ein entsprechend stärkeres Engagement beim Deutschlernen zeigen, neigen demnach stärker zum Self-Enhancement und somit zu einer Überschätzung ihrer Kompetenz als jugendliche Flüchtlinge, für die ihre Deutschkompetenz eine weniger wichtige Rolle spielt. Deshalb wird folgende Hypothese aufgestellt:

Hypothese 2b: Das Engagement beim Deutschlernen hat einen positiven Effekt auf die Selbsteinschätzung der Deutschkompetenz.

Ein weiterer Einfluss auf die Selbsteinschätzungen wird für die kognitiven Fähigkeiten, insbesondere die Fähigkeit zum schlussfolgernden Denken, vorhergesagt. Personen mit besser ausgeprägter Fähigkeit zum schlussfolgernden Denken haben mehr kognitive Kapazitäten und können

Aufgaben effizienter lösen, sodass sie weniger Ressourcen dabei verbrauchen als Personen mit niedriger ausgeprägter Fähigkeit zum schlussfolgernden Denken. Betrachtet man die komplexen Anforderungen im Selbsteinschätzungsprozess, steigern größere kognitive Kapazitäten und besser ausgeprägte Fähigkeiten im schlussfolgernden Denken die Wahrscheinlichkeit, dass diese erfolgreich gemeistert werden, während geringere kognitive Kapazitäten und Fähigkeiten im schlussfolgernden Denken eine Übersteigung der kognitiven Kapazität wahrscheinlicher machen mit Leistungseinbußen bei den entsprechenden Verarbeitungsprozessen und einem ungenauen Selbsteinschätzungsresultat. Komplexe Anforderungen stellen die Utilization-Stufe des RAM, auf der verschiedene Informationen verarbeitet und integriert werden müssen sowie die Beantwortung einer Selbsteinschätzungsfrage in einer konkreten Befragungssituation, in der die Frage verstanden werden muss, verfügbare und idealerweise bereits verarbeitete und integrierte Informationen abgerufen werden müssen und daraus ein Urteil gebildet und eine dazu passende Antwortoption gewählt werden soll. Auch auf der Detection-Stufe des RAM kann eine kognitive Überlastung dazu führen, dass verfügbare Hinweise aufgrund mangelnder kognitiver Kapazität nicht wahrgenommen werden, was die Genauigkeit der Selbsteinschätzung beeinträchtigen kann (vgl. Abschnitte 1.6.2.1 und 1.7.1).

Es wird folglich angenommen, dass die Selbsteinschätzungen umso genauer ausfallen, je höher die Fähigkeit zum schlussfolgernden Denken von Teilnehmenden ausgeprägt ist. Mit der genaueren Selbsteinschätzung einhergehend wird angenommen, dass das Ausmaß an Überschätzung reduziert wird. Deshalb wird folgende Hypothese aufgestellt:

Hypothese 2c: Die Fähigkeit zum schlussfolgernden Denken hat einen negativen Effekt auf die Selbsteinschätzung der Deutschkompetenz.

Zuletzt wird argumentiert, dass sich die Teilnahme an Deutschunterricht und einem Deutschtest auf die Selbsteinschätzung der Deutschkompetenz auswirkt. Direkte und indirekte soziale Rückmeldungen sind Quellen selbstbezogenen Wissens (vgl. Abschnitt 1.2.1). Dabei kann Feedback zusätzliche (Quantität), hochwertige (Qualität) Informationen zu einer genauen Selbsteinschätzung beitragen (vgl. Abschnitt 1.6.2.4). Nach dem Dunning-Kruger-Effekt überschätzen Personen ihre Leistung, wenn sie ihre Fehler nicht als solche erkennen können (vgl. Abschnitt 1.6.2.1). Diese Wissenslücke könnte Feedback von anderen Personen mit höherer Kompetenz schließen, indem diese auf Fehler aufmerksam machen. Direktes Feedback zur Richtigkeit der Lösung von Aufgaben oder Fehlern wird in der Regel in Unterrichtsettings gegeben. Inwiefern sich Feedback im Rahmen einer Teilnahme an Unterricht in einer Kompetenz auf die Genauigkeit von Selbsteinschätzungen dieser Kompetenz auswirkt, ist jedoch nach meinem Wissen nicht empirisch untersucht. Im Rahmen von Deutschunterricht könnten jugendliche Geflüchtete auf Fehler aufmerksam gemacht werden, welche sie sonst nicht erkennen würden, und somit könnte die

Überschätzung der eigenen Deutschkompetenz reduziert werden. Deshalb wird zum einen untersucht, ob die Teilnahme an Deutschunterricht die Selbsteinschätzung der Deutschkompetenz inkrementell zur objektiven Kompetenz beeinflusst. Darüber hinaus kann Feedback aus objektiven Tests eine zusätzliche Informationsquelle für genaue Einschätzungen sein. Testergebnisse haben den Vorteil, dass sie objektiver sind als soziale Rückmeldungen und häufig bieten sie eine Referenz, an der man die eigene Leistung messen kann. Dies könnte das Ergebnis des fehleranfälligen Prozesses der Integration von wahrgenommenen, verfügbaren Informationen über eigene Kompetenzen und Kompetenzen anderer Personen ergänzen. Deshalb wird zum anderen untersucht, ob die Teilnahme an einem Deutschtest die Selbsteinschätzung der Deutschkompetenz beeinflusst. Unter der Annahme, dass Jugendliche nur an einem Deutschtest teilgenommen haben, wenn sie auch einen Deutschkurs besucht haben, werden folgende Hypothesen aufgestellt:

Hypothese 2d: *Jugendliche, die an einem Deutschkurs teilgenommen haben, überschätzen ihre Deutschkompetenz weniger als Jugendliche, die an keinem Deutschkurs teilgenommen haben.*

Hypothese 2e: *Jugendliche, die sowohl an einem Deutschkurs als auch an einem Deutschtest teilgenommen haben, überschätzen ihre Deutschkompetenz weniger als Jugendliche, die an keinem Deutschtest, aber ggf. an einem Deutschkurs teilgenommen haben.*

2.3 Hypothesen zu verschiedenen Selbsteinschätzungsitems

Die Hypothesen in diesem Abschnitt beziehen sich auf die Fragestellung zu den Vor- und Nachteilen verschiedener Arten von Selbsteinschätzungsitems. In die Befragungen der ReGES-Studie konnten verschiedene Items zur Selbsteinschätzung der Deutschkompetenzen aufgenommen werden. Um die Vergleichbarkeit mit Stichproben des NEPS zu gewährleisten, wurden die dort verwendeten Items mit geringfügigen Änderungen übernommen (vgl. Nationales Bildungspanel (NEPS), 2013). In der ReGES-Studie lautet das Standard-Item zum Verstehen der deutschen Sprache „Wie gut verstehen Sie Deutsch?“ mit den Antwortkategorien *1 sehr gut*, *2 eher gut*, *3 eher schlecht*, *4 sehr schlecht* und *5 gar nicht* (s. *Abbildung 2* in Abschnitt 3.3.1.1). Das Item hat den Vorteil, dass es sehr effizient und einfach verständlich ist. In der Startkohorte 4 (SC4) des NEPS hat die Mehrheit der Teilnehmenden mit Migrationshintergrund jedoch die beiden positiven Antwortkategorien *eher gut* und *sehr gut* gewählt und Analysen zur Validität haben ergeben, dass die Items keine genauen Einschätzungen der Sprachkompetenz liefern können (Edele et al., 2015; s. Kapitel 1.9). Da wenige empirische Befunde für besser geeignete Selbsteinschätzungsitems vorlagen, wurden drei weitere Arten von Selbsteinschätzungsitems eingesetzt, sodass nun ein direkter

Vergleich der Items anhand der Daten der ReGES-Studie durchgeführt werden kann. Unter Berücksichtigung der Implikationen für die Gestaltung von Selbsteinschätzungsitems aus Abschnitt 1.7.2 wird im Weiteren dargelegt, welche weiteren Items in der ReGES-Studie zur Selbsteinschätzung der Deutschkompetenzen eingesetzt wurden, und es werden Hypothesen zu deren Vorteilen aufgestellt.

2.3.1 Schieberegler-Item

Verschiedene Möglichkeiten der Gestaltung der Selbsteinschätzungsitems könnten dazu beitragen, die Qualität der gewonnenen Daten zu verbessern. Betrachtet man die Ergebnisse der Selbsteinschätzungen von Deutschkompetenzen der Jugendlichen mit Migrationshintergrund in der Startkohorte 4 des NEPS, fällt der Deckeneffekt auf und dass die Teilnehmenden Kategorien mit positiver Bedeutung (*sehr gut* und *eher gut*) bevorzugen. Da die Antwortskala nur zwei Kategorien mit positiver Bedeutung beinhaltet, könnte sie nicht differenziert genug sein, um möglicherweise differenziertere Selbsteinschätzungen in diesem Bereich zu erfassen und die Antwortskala deckt womöglich nicht das gesamte Messkontinuum ab. Manchen Jugendlichen könnte eine noch bessere Kategorie als *sehr gut* fehlen, anderen eine mittlere Kategorie zwischen *eher schlecht* und *eher gut* (vgl. Abschnitt 1.7.2). Um mehr positive Antwortkategorien anbieten zu können und eine differenziertere Erfassung der Selbsteinschätzungen zu ermöglichen, wurde der Empfehlung von Leung (2011) entsprechend eine zehnstufige Antwortskala mit gelabelten Endpunkten von *1 sehr schlecht* bis *10 sehr gut* mit einer gesondert dargestellten Antwortoption *0 gar nicht* eingesetzt. Die Skala wurde als Schieberegler dargestellt. Die Frage blieb unverändert (s. *Abbildung 3* in Abschnitt 3.3.1.2). Eine solche Ratingskala mit mehr Antwortoptionen kann die Einschätzungen der Teilnehmenden differenzierter erfassen und die Schiefe der Verteilung reduzieren (vgl. Abschnitt 1.7.2). Wegen der Möglichkeit zur differenzierteren Erfassung wird erwartet, dass die Diskrimination der Skala besser ist als die der Standard-Items. Demnach werden folgende Hypothesen aufgestellt:

Hypothese 3a: *Die Verteilung der Antworten auf das Schieberegler-Item ist weniger schief als die Verteilung der Antworten auf das Standard-Item.*

Hypothese 3b: *Das Schieberegler-Item korreliert höher mit den objektiven Kompetenzmaßen als das Standard-Item.*

2.3.2 Vergleich-Item

Ein weiteres Problem, das die Korrelation zwischen Selbsteinschätzungen und Kompetenzscores beeinträchtigt, ist, dass Teilnehmende und auch Forschende unterschiedliche

Referenzrahmen bei der Selbsteinschätzung annehmen (vgl. Abschnitte 1.2.4 und 1.2.5). Um die angenommenen Referenzrahmen aneinander anzugleichen, empfiehlt es sich, einen Referenzrahmen in der Fragestellung vorzugeben und die Labels der Antwortskala relativ zu formulieren (vgl. Abschnitt 1.7.2), wie es z.B. in einer Elternbefragung der Startkohorte 2 des NEPS gemacht wurde, in der die Eltern die sprachlichen Fähigkeiten ihres Kindes im Vergleich zu anderen Kindern gleichen Alters einschätzen sollten, auf einer fünfstufigen gelabelten Antwortskala von *viel schlechter* bis *viel besser* (Leibniz-Institut für Bildungsverläufe e.V [LIfBi], 2018, S. 78). Eine Schwierigkeit besteht in der Auswahl einer geeigneten Referenz für die jugendlichen Flüchtlinge. Wenn Teilnehmende nicht beispielsweise alle gemeinsam einen Sprachkurs besuchen, gibt es keine Referenz, die für alle Teilnehmenden gleich ist. Ein Ziel der Selbsteinschätzungsfrage ist es, die Deutschkompetenzen der Teilnehmenden so abzubilden, dass sie zwischen den Teilnehmenden hinsichtlich der Kompetenz differenzieren. Naheliegend wäre es deshalb, andere jugendliche Geflüchtete als Referenz vorzugeben. Jedoch ist zu erwarten, dass das durchschnittliche Sprachniveau des direkten Umfelds der Geflüchteten stark variiert, u.a. in Abhängigkeit von der Aufenthaltsdauer der Individuen und somit ebenfalls einen ungleichen Referenzrahmen bietet. Einen auch im Sinne des Shifting-Standards-Effekts objektiveren Anhaltspunkt bietet ein klar definiertes Sprachniveau als Referenz. Als solches sollte das Muttersprachniveau allen Teilnehmenden vertraut sein, da sie einerseits selbst mindestens eine Sprache auf Muttersprachniveau beherrschen und andererseits im Alltag in Deutschland sowie im Schulkontext mit Personen in Kontakt sind, deren Muttersprache Deutsch ist. Entsprechend wurden Personen, deren Muttersprache Deutsch ist, als Vergleich herangezogen. Dies hat daneben den Vorteil, dass für Personen, die im Vergleich zu Muttersprachlern erst seit sehr kurzer Zeit die deutsche Sprache lernen, ein möglicher Deckeneffekt der Selbsteinschätzung reduziert werden sollte. Problematisch könnte bei der Vorgabe von Muttersprachlern als Referenz allerdings sein, dass Personen es bevorzugen, sich mit anderen zu vergleichen, die ähnliche Fähigkeiten besitzen (Festinger, 1954) und es Personen mit noch schlecht ausgeprägten Sprachkompetenzen schwer fallen könnte, sich mit Personen auf Muttersprachniveau zu vergleichen. Trotz dieser Einschränkungen wird angenommen, dass die Vorgabe eines Referenzrahmens einen positiven Einfluss auf die Diskrimination der Selbsteinschätzungen hat. Das Vergleich-Item zum Verstehen der deutschen Sprache lautet: „Bitte vergleichen Sie sich nun mit Gleichaltrigen, deren Muttersprache Deutsch ist. Wie gut verstehen Sie Deutsch?“ Es konnten diese Antwortoptionen gewählt werden: *1 ich spreche gar kein Deutsch, 2 ich spreche Deutsch viel schlechter als ein Deutscher, 3 ich spreche Deutsch schlechter als ein Deutscher, 4 ich spreche Deutsch fast genauso gut wie ein Deutscher, 5 ich spreche Deutsch genauso gut wie ein Deutscher* (s. *Abbildung 4* in Abschnitt 3.3.1.3). Zum Deckeneffekt und zur Diskrimination des Vergleich-Items werden folgende Hypothesen aufgestellt.

Hypothese 4a: *Die Verteilung der Antworten auf das Vergleich-Item zeigt keinen Deckeneffekt.*

Hypothese 4b: *Das Vergleich-Item korreliert höher mit den objektiven Kompetenzmaßen als das Standard-Item.*

2.3.3 Can-Do-Statements

Die beschriebene Bevorzugung der positiven Antwortkategorien der Jugendlichen mit Migrationshintergrund deutet des Weiteren darauf hin, dass die Teilnehmenden bei der Beantwortung der Frage eher zum Self-Enhancement als zum Self-Assessment neigten. Self-Enhancement wird begünstigt, wenn die Frage uneindeutig formuliert ist und Interpretationsspielraum bietet (vgl. Kapitel 1.3). Dies ist bei dem Standard-Item zum Verstehen der deutschen Sprache der Fall, denn auch die linguistische Dimension Verstehen umfasst verschiedene Fähigkeitsbereiche, wie z.B. Grammatik und Wortschatz, zu denen Einschätzungen in der Fragebogensituation kaum vollumfänglich abgerufen und gewichtet werden können. Die Teilnehmenden haben die Möglichkeit, die Fragen wohlwollend auszulegen, indem sie Informationen zu den Fähigkeitsbereichen abrufen, in denen sie gut abschneiden (vgl. Kapitel 1.3).

Zudem wird angenommen, dass es für Teilnehmende einfacher ist, Fragen zu beantworten, die sich auf bekannte Situationen beziehen, in denen sie vielfältige Erfahrungen gesammelt haben (vgl. Abschnitt 1.7.2). In verschiedenen Kontexten, insbesondere durch den Gemeinsamen Europäischen Referenzrahmen (GER; Council of Europe, 2001), wurden sogenannte Can-Do-Statements verwendet, um Sprachkompetenzen von Personen zu erfassen. Die Can-Do-Statements beziehen sich auf konkrete alltägliche Fähigkeiten, welche gleichzeitig eindeutiger definierte Fähigkeiten darstellen als es bei den bisher beschriebenen Items der Fall ist. Ein Can-Do-Statement des DIALANG-Projekts lautet übersetzt z.B. „Ich kann Fragen und Anweisungen verstehen und kurzen, einfachen Anweisungen folgen“ (Council of Europe, 2001, S. 233). Jedes Statement kann einem der Kompetenzniveaus des GER, die von A1 bis C2 reichen, zugeordnet werden. Solche alltagsnahen Statements, die nur mit ja oder nein beantwortet werden müssen, könnten die Schwierigkeit der Selbsteinschätzung reduzieren, da die Fähigkeit, zu der Informationen abgerufen werden müssen, weniger umfangreich, weniger abstrakt und klarer definiert ist. Die Integration der Informationen zu einer Einschätzung des Fähigkeitsniveaus kann dann anhand der Angaben durch die Forschenden vorgenommen werden. Weiterhin könnte die Alltagsnähe die Motivation der Teilnehmenden bei der Beantwortung der Frage steigern. Es wird erwartet, dass die höhere Motivation auf Seiten der Teilnehmenden sowie die geringere Aufgabenschwierigkeit zu einer gründlicheren Verarbeitung bei der Beantwortung der Fragebogenitems und somit zu genaueren Selbsteinschätzungen führen (vgl. Abschnitt 1.7.1).

Aus diesen Gründen wurden Items des auf den Can-Do-Statements des GER basierenden „XS-Tests“ (Deutscher Volkshochschul-Verband e.V., 2011) ausgewählt, teilweise angepasst und zu einer Selbsteinschätzungsskala zusammengefügt. Der XS-Test dient der Selbstzuordnung zu Sprachkursen des richtigen Kompetenzniveaus. Er beinhaltet fünf einfach verständliche alltagsbezogene Can-Do-Statements für jedes der sechs Kompetenzlevel des GER. Um die Anzahl der Items auf ein effizienteres Maß zu reduzieren, wurden 14 der 30 Can-Do-Statements ausgewählt und dabei darauf geachtet, dass die gewählten Items einfach formuliert sind und sich möglichst auf Situationen beziehen, von denen erwartet wurde, dass die Teilnehmenden diese in deutschsprachigen Kontexten erlebt haben. Außerdem wurde darauf geachtet, dass sich die Items über alle Kompetenzniveaus und die Domänen Verstehen, Sprechen, Lesen und Schreiben verteilen. Die 14 Items sind in *Abbildung 5* in Abschnitt 3.3.1.4 nachzulesen. Dem Verstehen und Sprechen der deutschen Sprache sind acht Items zuzuordnen. Die übergeordnete Aufgabe lautete: „Bitte geben Sie alles an, was Sie auf Deutsch können.“ Die Teilnehmenden konnten die einzelnen Can-Do-Statements auswählen oder nicht auswählen. Für den Fall, dass jemand noch gar nichts davon konnte, gab es die Antwortoption *15 nichts davon*. Wegen der einfachen, eindeutigen und alltagsnahen Formulierungen, wird angenommen, dass die Teilnehmenden die Items genauer beantworten und auf den Angaben basierende Kompetenzbeurteilungen genauer sind als Selbsteinschätzungen mittels Standard-Item. Es wird folgende Hypothese aufgestellt.

Hypothese 5a: Die aus den Can-Do-Statements abgeleiteten Kompetenzbeurteilungen korrelieren stärker mit den objektiven Kompetenzmaßen als das Standard-Item.

Can-Do-Statements haben außerdem den Vorteil, dass sie zu einem großen Teil auf Informationen zurückgreifen, die über Selbstwahrnehmungen gewonnen wurden und weniger auf sozialen Vergleichen und Referenzrahmen beruhen (vgl. Kapitel 1.2). Deshalb können sie voraussichtlich besser Veränderungen der Sprachkompetenzen über die Zeit erfassen als Items, die stärker durch Referenzrahmen beeinflusst werden, denn Referenzrahmen verändern sich ebenfalls mit der Zeit und können somit keinen stabilen Standard bieten. Wenn sich also z.B. die Kompetenz einer Person verbessert hat, aber ebenso die Kompetenz der Personen im direkten Umfeld besser geworden ist, würde sich die Person auf den Can-Do-Items zu einem zweiten Messzeitpunkt voraussichtlich besser einschätzen als zum ersten Messzeitpunkt. Bei Items, die Vergleiche mit anderen implizieren, würde die Verbesserung der Kompetenz zum zweiten Messzeitpunkt hingegen vermutlich nicht angemessen durch die Selbsteinschätzung erfasst werden. Problematisch bei der Erfassung von Entwicklungen ist jedoch auch bei den Can-Do-Statements, dass das Ausmaß an Überschätzung mit steigender Kompetenz entsprechend dem Dunning-Kruger-Effekt sinken sollte (vgl. Abschnitt 1.6.2.1). Verbesserungen der Sprachkompetenz könnten also auch durch die Can-Do-Statements unterschätzt werden. Da jedoch zumindest angenommen werden kann, dass die

Referenzrahmen-Effekte reduziert werden, wird erwartet, dass die Can-Do-Statements die Veränderungen der Deutschkompetenz zwischen zwei Messzeitpunkten besser erfassen können als die Standard-Items, wozu die entsprechende Hypothese aufgestellt wird.

Hypothese 5b: Die Can-Do-Statements können Veränderungen der Deutschkompetenz zwischen zwei Messzeitpunkten besser erfassen als die Standard-Items.

3 Methode

Im ersten Teil dieses Kapitels wird über die verwendeten Daten und die Auswahl der Stichproben informiert und die Stichproben werden beschrieben. Ein Abschnitt widmet sich dabei der Untersuchung der Selektivität der ausgewählten Stichproben. Im zweiten Teil des Kapitels wird auf den Ablauf in beiden Erhebungswellen eingegangen und im dritten Teil des Kapitels werden die Erhebungsinstrumente präsentiert und erläutert, wie aus den Antworten der Teilnehmenden die in den Analysen verwendeten Variablen gebildet wurden.

3.1 Daten und Stichprobe

Für die Analysen dieser Arbeit wurden die im Scientific-Use-File Version 3.0.0 veröffentlichten Daten der Jugendlichen-Kohorte der Studie *Refugees in the German Educational System* (ReGES-Studie) verwendet.[1] Will, Homuth, von Maurice und Roßbach (2021) geben einen Überblick über die Studie und deren Forschungspotenzial. Im Folgenden werden zuerst allgemeine Informationen zur ReGES-Studie und den ReGES-Daten zusammengefasst, anschließend wird beschrieben, welche Daten der ReGES-Studie für die Analysen herangezogen wurden und welche Fälle in die Analysestichproben einbezogen wurden. Die Analysestichproben werden beschrieben und es wird auf die Selektivität der betrachteten Stichproben eingegangen.

3.1.1 Informationen zu den Daten der ReGES-Studie

Die ReGES-Studie ist eine Panelstudie mit dem Fokus auf geflüchteten Kindern und Jugendlichen. Mithilfe der Daten sollen die Teilnahme der Kinder und Jugendlichen am deutschen Bildungssystem beschrieben sowie für die Integration förderliche und hinderliche Faktoren identifiziert werden können. In die Jugendlichen-Kohorte wurden Jugendliche aufgenommen, die im Oktober 2017 14 bis 16 Jahre alt waren und zum Zeitpunkt der ersten Erhebung die untere Sekundarstufe (fünfte bis zehnte Klasse) des allgemeinbildenden Schulsystems besuchten. Die Studie bestand aus sieben Erhebungswellen. Im jährlichen Abstand wurden drei computergestützte persönliche Interviews durchgeführt (CAPI und CASI; Wellen 1, 4 und 7). Weiterhin gab es ein Telefoninterview (CATI; Welle 3) und drei Onlinebefragungen (CAWI; Wellen 2, 5 und 6). Die Jugendlichen wurden in jeder Welle selbst befragt. In der ersten Welle wurden die Eltern der Jugendlichen

[1] Diese Arbeit nutzt Daten der Studie "Refugees in the German Educational System" (ReGES): Refugee Cohort 2 – Jugendliche, doi:10.5157/ReGES:RC2:SUF:3.0.0. Die Studie wurde von Juli 2016 bis Dezember 2021 unter der Fördernummer FLUCHT03 vom Bundesministerium für Bildung und Forschung (BMBF) gefördert und am Leibniz-Institut für Bildungsverläufe (LIfBi) verantwortet.

optional zusätzlich befragt. In den Befragungen ging es um verschiedene Inhalte, wie z.B. die Bildungsbiographie der Jugendlichen, die Bildung der Eltern, Schulleistungen, Bildungsaspirationen, Deutschkompetenzen und verschiedene Aspekte des Deutscherwerbs, die Rückkehrorientierung, Religion, schul- oder ausbildungsbezogene motivationale Aspekte, wahrgenommene Unterstützung von Geflüchteten und wahrgenommene Diskriminierung, den Gesundheitszustand, Resilienz und Zufriedenheit (s. a. Will et al., 2021). Die Interviews in der ersten und siebten Welle wurden durch Testungen der Deutschkompetenzen und kognitiver Grundfähigkeiten ergänzt. Darüber hinaus wurden zu verschiedenen Erhebungszeitpunkten auch institutionelle und regionale Kontextpersonen schriftlich befragt. Dazu gehören Schulleitungen, Lehrerinnen und Lehrer und Mitarbeitende der Gemeinden und Gemeinschaftsunterkünfte (Will et al., 2021).

Die Stichprobe der ReGES-Studie wurde in fünf deutschen Bundesländern gezogen: Bayern, Hamburg, Nordrhein-Westfalen, Rheinland-Pfalz und Sachsen (detailliertere Informationen zur Stichprobenziehung s. Steinhauer et al., 2019). Um in die Stichprobe aufgenommen zu werden, mussten die Jugendlichen neben den oben genannten Einschränkungen bezüglich des Alters und des Schulbesuchs folgende Kriterien erfüllen: Sie mussten die Befragung in einer der acht Interviewsprachen (Deutsch, Arabisch, Englisch, Farsi, Französisch, Kurmandschi, Paschtu, Tigrinya) durchführen können. Sie mussten nach dem 1. Januar 2014 als Flüchtling nach Deutschland gekommen sein und seit mindestens drei Monaten in Deutschland leben. Weiterhin musste eine erziehungsberechtigte Person mit im Haushalt leben. Insgesamt wurden in der ersten Erhebungswelle mit 2 415 Jugendlichen gültige Interviews durchgeführt (z.B. Steinhauer et al., 2019). Für die Kompetenztestung und die weiteren Befragungswellen wurden die angebotenen Sprachen auf Arabisch, Deutsch, Englisch und Kurmandschi begrenzt. Um an der Kompetenztestung, der Befragung der institutionellen Kontextpersonen und an den Befragungen der folgenden Wellen teilnehmen zu können, mussten die Jugendlichen ihr Einverständnis geben, erneut kontaktiert zu werden und angeben, eine der vier *Panelsprachen* zum Verständnis und zur Beantwortung der Befragung ausreichend zu beherrschen. Diese beiden Kriterien erfüllten 2 267 Jugendliche (94%), wobei nur sechs Fälle ausgeschlossen wurden, die zwar panelbereit waren, jedoch keine der Panelsprachen ausreichend beherrschten (Becker et al., 2021). Während die Teilnahme an der Kompetenztestung nur für die 2 267 panelfähigen Jugendlichen möglich war, konnte die Kompetenztestung unabhängig von der Panelteilnahme bereits im Screening ausgeschlossen und auch kurzfristig beim Übergang der Befragung zum Start der Kompetenztestung verweigert werden. Für 2 016 Jugendliche wurde die Kompetenztestung in der ersten Erhebungswelle tatsächlich administriert, wobei aus verschiedenen Gründen, wie z.B. Testabbrüchen, Kompetenzwerte auch für einige dieser Jugendlichen fehlten (s. Abschnitt 3.3.2).

3.1.2 Verwendete Daten und Analysestichproben

Zur Beantwortung der Fragestellung wurden zum einen Daten der Befragung und Kompetenztestung der Jugendlichen der ersten Erhebungswelle herangezogen. Als Bedingung zur Aufnahme in die für diese Arbeit herangezogene Stichprobe der ersten Erhebungswelle musste mindestens eine gültige Antwort auf das Standard-Selbsteinschätzungsitem zum Verstehen der deutschen Sprache vorliegen sowie mindestens ein gültiger Deutschkompetenzscore für einen der beiden eingesetzten Deutschkompetenztests. Diese Bedingungen erfüllten $N = 1\,877$ Fälle und bildeten somit die Analysestichprobe der ersten Erhebungswelle für diese Arbeit, welche ich im Folgenden *Analysestichprobe 1* nenne. Von den 139 Fällen, die an der Kompetenztestung teilgenommen hatten, aber nicht in die Analysestichprobe 1 aufgenommen wurden, lag für 4 Fälle keine gültige Antwort auf das Standard-Selbsteinschätzungsitem zum Verstehen aber ein Deutschkompetenzscore vor, für 131 Fälle lag für keinen der beiden Deutschkompetenzscores ein gültiger Wert vor, aber eine gültige Antwort auf das Standard-Selbsteinschätzungsitem. Für 4 Fälle lag weder eine gültige Antwort auf das Standard-Selbsteinschätzungsitem zum Verstehen vor noch ein gültiger Deutschkompetenzscore. Auf die Gründe für die fehlenden Werte der Kompetenztests wird in Abschnitt 3.3.2 genauer eingegangen (s. a. Obry et al., 2021).

In den Analysen dieser Arbeit wurden neben den Daten der Befragung und der Kompetenztestung der Jugendlichen der ersten Welle auch Angaben der Lehrerinnen und Lehrer der ersten Welle zur Mathematiknote der Jugendlichen verwendet. Zur Befragung der institutionellen Kontextpersonen mussten die Jugendlichen und deren Eltern ihr schriftliches Einverständnis geben und Angaben zur Schule und Klasse machen, welche daraufhin ermittelt wurden. In 1 615 Fällen der Gesamtstichprobe der Kohorte der Jugendlichen der ReGES-Studie (vgl. Becker et al., 2021) und in 1 361 Fällen der Analysestichprobe 1 konnten die Einrichtungen für die Befragungen ermittelt werden. Für 330 Jugendliche der Analysestichprobe 1 lag in der ersten Erhebungswelle ein ausgefüllter Schülerfragebogen vor.

Weiterhin habe ich Analysen dieser Arbeit auch anhand der Daten der Befragung und Kompetenztestung der Jugendlichen der siebten Erhebungswelle durchgeführt. Wie auch für die Analysestichprobe 1 war die Voraussetzung für die Aufnahme in die Analysestichprobe der siebten Erhebungswelle, im Folgenden *Analysestichprobe 2* genannt, dass in dieser Welle mindestens eine gültige Antwort auf das Standard-Selbsteinschätzungsitem zum Verstehen der deutschen Sprache vorlag sowie mindestens ein gültiger Deutschkompetenzscore für einen der beiden eingesetzten Deutschkompetenztests. Da die siebte Erhebungswelle für den Zeitraum zwischen Februar und Mai 2020 geplant war und in Form von persönlichen Interviews stattfand, musste die Feldphase am 16. März 2020 aufgrund der Covid-19 Pandemie abgebrochen werden, um sowohl Interviewende als auch Teilnehmende vor einer Ansteckung mit dem Coronavirus zu schützen. Zu diesem

Zeitpunkt waren bereits 836 Befragungen mit gültigen Daten der Stichprobe der Jugendlichen durchgeführt worden. Die Befragungen der restlichen Stichprobe wurden nach kurzer Umstellung zwar in Form von Telefoninterviews wiederaufgenommen, dies hatte jedoch den Nachteil, dass für diese Fälle keine Kompetenztestung durchgeführt werden konnte (Will et al., 2020). Insgesamt wurde bei 794 Jugendlichen die Kompetenztestung in der siebten Erhebungswelle administriert. Für 778 dieser Fälle lag ein gültiger Summenwert für einen der beiden Deutschkompetenztests vor. Ein gültiger Wert für das Standard-Selbsteinschätzungsitem lag für alle dieser Fälle vor, sodass die 778 Fälle die Analysestichprobe 2 bildeten. 72 der Fälle der Analysestichprobe 2 waren nicht Teil der Analysestichprobe 1, dies lag in allen Fällen daran, dass in der ersten Erhebungswelle kein gültiger Deutschkompetenzscore vorlag.

3.1.3 Beschreibung der Analysestichproben

Im Folgenden wird zuerst die Analysestichprobe 1 und anschließend die Analysestichprobe 2 beschrieben. Dabei wird auf demographische Angaben wie Alter, Geschlecht, Staatsangehörigkeit und Muttersprache eingegangen sowie auf die Aufenthaltsdauer in Deutschland, auf die Bundesländer, in denen die Befragten lebten, auf die Unterkünfte, in der die Befragten in Deutschland lebten und auf die Dauer des Schulbesuchs in Deutschland. Für die Analysestichprobe 2 wird zudem darauf eingegangen, wie viele Befragte sich zum Zeitpunkt der siebten Erhebungswelle noch im allgemeinbildenden Schulsystem oder in verschiedenen weiteren Ausbildungs- oder Erwerbssituationen befanden.

3.1.3.1 Analysestichprobe 1

Die $N = 1\,877$ jugendlichen Flüchtlinge der Analysestichprobe 1 waren zum Zeitpunkt der Befragung der ersten Erhebungswelle zwischen 14.3 und 17.6 Jahre alt ($M = 16.0$ Jahre, $SD = 0.9$ Jahre). Der Anteil männlicher Teilnehmender betrug 56% und 44% der Teilnehmenden waren weiblichen Geschlechts. Eine syrische Staatsangehörigkeit hatten 71%, eine irakische Staatsangehörigkeit hatten 13%, eine afghanische Staatsangehörigkeit hatten 6% der Teilnehmenden und 7% hatten eine andere oder keine Staatsangehörigkeit. Für 2% der Fälle wurde keine Angabe zur Staatsangehörigkeit gemacht. Als Muttersprache gaben 77% der Jugendlichen Arabisch, 8% Persisch, 29% Kurdisch, 11% Englisch und 22% Deutsch an. Weitere Sprachen gaben jeweils unter 5% der Teilnehmenden an. Die Muttersprache war definiert als die Sprache, die die Teilnehmenden in den ersten drei Lebensjahren in ihrer Familie gelernt haben. Es war möglich, mehrere Antworten auszuwählen. Eine Muttersprache gaben 66% an, 18% gaben zwei Muttersprachen an und 15% gaben drei oder mehr Muttersprachen an. Da die Teilnehmenden nur in die Studie aufgenommen wurden, wenn sie nach dem 1. Januar 2014 als Flüchtlinge nach Deutschland gekommen

waren und zu diesem Zeitpunkt mindestens 10 Jahre alt waren, ist zu vermuten, dass einige Jugendliche die Frage falsch verstanden und beispielsweise alle Sprachen, die sie sprechen, ausgewählt haben. Anders ist der Anteil von 22% der Teilnehmenden, die Deutsch als Muttersprache angegeben haben, kaum zu erklären.

Im Mittel lebten die Teilnehmenden seit 2.5 Jahren in Deutschland (SD = 0.8 Jahre, Min = 0.2 Jahre, Max = 4.4 Jahre). Die Teilnehmenden lebten zum Zeitpunkt der ersten Befragung entsprechend der Stichprobendefinition in fünf deutschen Bundesländern. In Hamburg lebten 5% der Teilnehmenden, 65% lebten in Nordrhein-Westfalen, 9% in Rheinland-Pfalz, 12% in Bayern und 9% in Sachsen. In einer privaten Unterkunft lebten zum Zeitpunkt der ersten Befragung 91% der Jugendlichen und 9% lebten in einer Gemeinschaftsunterkunft. Eine Schule in Deutschland besuchten die 1 849 Jugendlichen, für die diese Angabe vorlag, seit durchschnittlich 1.9 Jahren (SD = 0.8 Jahre, Min = 0 Jahre, Max = 4.4 Jahre).

3.1.3.2 Analysestichprobe 2

Die N = 778 jugendlichen Flüchtlinge der Analysestichprobe 2 waren zum Zeitpunkt der siebten Erhebungswelle im Durchschnitt 17.8 Jahre alt (SD = 0.9 Jahre, Min = 14.6 Jahre, Max = 20.2 Jahre). In der vierten und siebten Erhebungswelle bestand die Möglichkeit, das in vorherigen Wellen erfasste Geburtsdatum zu korrigieren, was die Bereiche im Altersrange erklärt, die aufgrund der Aufnahmekriterien in die Stichprobe der ersten Erhebungswelle nicht möglich wären. Ein im Vergleich zur ersten Erhebungswelle korrigiertes Geburtsdatum lag zum Zeitpunkt der siebten Erhebungswelle in 72 Fällen vor. Das anhand des Geburtsdatums berechnete Alter wurde in diesen Fällen maximal um 2.0 Jahre nach unten bis maximal um 2.3 Jahre nach oben korrigiert (M = –0.0, SD = 1.0). Männlichen Geschlechts waren 55% der Analysestichprobe 2 und 45% waren weiblichen Geschlechts. Auch für das Geschlecht war eine Änderung der Angabe aus vorherigen Wellen möglich. Eine Änderung des angegebenen Geschlechts wurde in 9 Fällen vorgenommen, wobei die Geschlechtsangabe in 2 Fällen von weiblich zu männlich und in 7 Fällen von männlich zu weiblich geändert wurde. Eine syrische Staatsangehörigkeit hatten 77% der Jugendlichen der Analysestichprobe 2, eine irakische Staatsangehörigkeit hatten 11%, eine afghanische Staatsangehörigkeit hatten 4% und 7% hatten eine andere oder keine Staatsangehörigkeit. In weniger als 1% der Fälle fehlte die Angabe. Die Muttersprache wurde ausschließlich in der ersten Befragung erhoben, wobei wie oben beschrieben die Möglichkeit bestand, mehrere Muttersprachen zu wählen. Von den Fällen, die zur Analysestichprobe 2 gehörten, haben 83% Arabisch als Muttersprache angegeben, 5% haben Persisch als Muttersprache angegeben, 27% haben Kurdisch als Muttersprache angegeben, 15% haben Englisch als Muttersprache angegeben, 26% haben Deutsch als Muttersprache angegeben und alle weiteren Sprachen wurden von einem geringeren Anteil der Teilnehmenden als

Muttersprache angegeben. Wie oben bereits erläutert ist zu vermuten, dass manche Teilnehmenden auch Sprachen angegeben haben, die nicht ihre Muttersprache waren. Eine Muttersprache haben 63% der Befragten der Analysestichprobe 2 angegeben, 18% haben zwei Muttersprachen angegeben und 19% haben drei oder mehr Muttersprachen angegeben.

Die Jugendlichen lebten zum Zeitpunkt der siebten Erhebungswelle im Durchschnitt seit 4.3 Jahren in Deutschland ($SD = 0.8$ Jahre, $Min = 2.1$ Jahre, $Max = 6.2$ Jahre). Zu diesem Zeitpunkt lebten 3% der Befragten in Hamburg, 76% in Nordrhein-Westfalen, 11% in Rheinland-Pfalz, 5% in Bayern und 6% in Sachsen. Außerdem wohnten zu diesem Zeitpunkt 96% der Befragten der Analysestichprobe 2 in einer privaten Unterkunft, während 3% noch in einer Gemeinschaftsunterkunft lebten. Die entsprechende Angabe fehlte in weniger als 1% der Fälle. Während sich die Befragten zum Zeitpunkt der ersten Erhebungswelle entsprechend den Aufnahmevoraussetzungen der Studie noch alle im allgemeinbildenden Schulsystem befanden, besuchten zum Zeitpunkt der siebten Erhebungswelle noch 65% der Stichprobe eine allgemeinbildende Schule. In Ausbildung befanden sich 3% der Teilnehmenden, 25% machten eine Berufsvorbereitung, 1% war erwerbstätig und machte weder eine Ausbildung noch besuchten diese Jugendliche eine Schule und auf 5% traf keine der obigen Antwortmöglichkeiten zu und sie machten etwas anderes, wie z.B. einen Freiwilligendienst oder ein Praktikum zu absolvieren.

3.1.4 Selektivität der Analysestichproben

Um aufzuzeigen, inwiefern die Analysestichproben hinsichtlich verschiedener Merkmale selektiv waren, wurden die Analysestichproben jeweils mit der Stichprobe aller gültiger Fälle der ReGES-Daten der ersten Erhebungswelle verglichen, die nicht zu der jeweiligen Analysestichprobe gehörten. Insgesamt enthielten die ReGES-Daten der ersten Erhebungswelle $N = 2\,415$ gültige Fälle. Die Analysestichprobe 1 umfasste $N = 1\,877$ Fälle und wurde somit mit der entsprechenden Stichprobe der ausgeschlossenen $N = 538$ Fälle verglichen. Die Analysestichprobe 2 umfasste $N = 778$ Fälle und wurde somit mit der entsprechenden Stichprobe der ausgeschlossenen $N = 1\,637$ Fälle verglichen. Dazu sind deskriptive Statistiken der Selbsteinschätzungen der Deutschkompetenzen, der Kompetenztests der kognitiven Grundfähigkeiten und der Deutschkompetenzen, sowie die Variablen, die bereits zur Stichprobenbeschreibung herangezogen wurden, in *Tabelle 1* (metrische Variablen) und *Tabelle 2* (kategoriale Variablen) aufgelistet. Die Statistiken sind in den Tabellen jeweils getrennt für die beiden Analysestichproben und für die beiden Stichproben der ausgeschlossenen Fälle angegeben. Im Sinne der Vergleichbarkeit wurden ausschließlich Daten aus der

ersten Erhebungswelle verwendet[2], weshalb die Angaben für die Analysestichprobe 2 von der Stichprobenbeschreibung in Abschnitt 3.1.3.2 abweichen, wenn diese Informationen aus späteren Wellen beinhalten. Signifikanztests wurden zum Vergleich der Mittelwerte und Varianzen der beiden Analysestichproben jeweils mit der entsprechenden Stichprobe der ausgeschlossenen Fälle durchgeführt. Die Analysen wurden in R Version 4.0.3 durchgeführt (R Core Team, 2020). Die Signifikanz der Mittelwertsunterschiede wurde bei Varianzhomogenität mittels ungerichtetem *t*-Test für unabhängige Stichproben und bei Varianzheterogenität mittels ungerichtetem Welch-Test bestimmt. Die Varianzhomogenität bzw. die Signifikanz der Abweichung zwischen den Varianzen wurde mittels Levene-Tests geprüft. Dazu wurde das car-Paket Version 3.1-0 in R verwendet (Fox & Weisberg, 2019). Mittelwerts- oder Varianzunterschiede, die mindestens auf dem 5%-Niveau signifikant waren, wurden beim Mittelwert oder bei der Standardabweichung gekennzeichnet. Eine alpha-Korrektur wurde nicht durchgeführt, damit möglicherweise vorhandene Unterschiede unwahrscheinlicher übersehen werden. Weiterhin wurde bei signifikanten Mittelwertsunterschieden die Effektstärke (Cohen's *d*) berechnet. Dazu wurde das effectsize-Paket Version 0.7.0.5 in R verwendet (Ben-Shachar et al., 2020). Bei den Häufigkeitsverteilungen der kategorialen Variablen wurden keine Signifikanztests durchgeführt.

Tabelle 1. Selektivitätsanalyse: Vergleich der Analysestichproben mit den Stichproben ausgeschlossener Fälle (metrische Variablen)

	Analysestichprobe 1		**Ausgeschlossene Fälle 1**		**Analysestichprobe 2**		**Ausgeschlossene Fälle 2**	
	n	***M* (*SD*)**	***n***	***M* (*SD*)**	***n***	***M* (*SD*)**	***n***	***M* (*SD*)**
SE Verstehen – Standard	1 877	4.17 (0.59)	530	4.25* (0.65**)	778	4.18 (0.63)	1 629	4.19 (0.60)
SE Verstehen – Schieberegler	944	7.37 (1.59)	252	7.19 (1.86**)	375	7.32 (1.55)	821	7.34 (1.70**)
SE Verstehen – Vergleich	922	3.5 (0.76)	279	3.65** (0.82)	402	3.49 (0.75)	799	3.56 (0.79)
SE Verstehen und Sprechen – Can-Do-Statements Summenwert[a]	1 870	5.08 (1.79)	525	5.08 (1.81)	776	5.17 (1.65)	1 619	5.04 (1.86**)
DGCF-BZT Summenwert	1 620	32.35 (8.68)	74	26.23*** (13.33***)	581	32.69 (8.79)	1 113	31.77* (9.12)
DGCF-MAT Summenwert	1 506	6.41 (2.77)	37	5.32* (2.86)	544	6.82 (2.68)	999	6.14*** (2.80)

[2] Eine Ausnahme stellen die Informationen zur Staatsangehörigkeit dar. Die im Scientific-Use-File enthaltene generierte Variable beinhaltete auch in späteren Wellen eingeholte Informationen zur Staatsangehörigkeit, falls die Information zu dem Zeitpunkt noch gefehlt hatte.

	Analysestichprobe 1		Ausgeschlossene Fälle 1		Analysestichprobe 2		Ausgeschlossene Fälle 2	
	n	*M* (*SD*)	*n*	*M* (*SD*)	*n*	*M* (*SD*)	*n*	*M* (*SD*)
PPVT-4 Summenwert	1 391	89.73 (32.35)	-	-	498	89.38 (32.48)	895	89.9 (32.26)
TROG-D Summenwert	1 801	33.48 (5.87)	-	-	671	33.54 (5.70)	1 134	33.45 (5.96)
Alter	1 877	15.97 (0.86)	538	15.92 (0.87)	778	15.94 (0.84)	1 637	15.97 (0.88)
Aufenthaltsdauer (in Jahren)	1 877	2.46 (0.75)	538	2.46 (0.77)	778	2.44 (0.76)	1 637	2.47 (0.75)
Dauer Schulbesuch in Deutschland (in Jahren)	1 849	1.93 (0.82)	518	1.98 (0.84)	773	1.90 (0.82)	1 594	1.97 (0.82)

Anmerkungen: SE = Selbsteinschätzung der Deutschkompetenz. Alle Werte beziehen sich auf Daten der ersten Erhebungswelle. Nur für 2 bzw. 4 Fälle liegen Werte zur Deutschkompetenz (PPVT-4 bzw. TROG-D) in der Stichprobe der ausgeschlossenen Fälle der ersten Erhebungswelle vor, daher werden für diese keine Angaben gemacht. Angaben zur Signifikanz: Mittelwert oder Varianz weicht signifikant von dem Mittelwert oder der Varianz der Analysestichprobe der entsprechenden Welle ab. Zum Vergleich der Varianzen wurden Levene-Tests durchgeführt, zum Vergleich der Mittelwerte t-Tests für unabhängige Stichproben bei Varianzhomogenität und Welch-Tests bei Varianzheterogenität.

[a] Der Summenwert der Can-Do-Statements wurde aus den in *Abbildung 5* in Abschnitt 3.3.1.4 dargestellten Statements Nr. 1, 2, 4, 6, 7, 10 und 11 berechnet, vgl. Abschnitt 4.1.4.

* $p < .05$, ** $p < .01$, *** $p < .001$

Tabelle 2. Selektivitätsanalyse: Vergleich der Analysestichproben mit den Stichproben ausgeschlossener Fälle (kategoriale Variablen)

	Analysestichprobe 1		Ausgeschlossene Fälle 1		Analysestichprobe 2		Ausgeschlossene Fälle 2	
	n	%	*n*	%	*n*	%	*n*	%
Geschlecht								
männlich	1 042	56	288	54	434	56	896	55
weiblich	835	44	250	46	344	44	741	45
Herkunft								
Syrien	1 334	71	274	51	602	77	1 006	61
Irak	250	13	63	12	87	11	226	14
Afghanistan	119	6	111	21	33	4	197	12
Sonstige	135	7	43	8	54	7	124	8
keine Angabe	39	2	47	9	2	0	84	5
Muttersprache								
Arabisch	1 436	77	329	61	642	83	1 123	69
Persisch	145	8	126	23	40	5	231	14
Kurdisch	549	29	138	26	207	27	480	29
Englisch	213	11	37	7	116	15	134	8
Deutsch	412	22	95	18	200	26	307	19

	Analysestichprobe 1		Ausgeschlossene Fälle 1		Analysestichprobe 2		Ausgeschlossene Fälle 2	
	n	%	*n*	%	*n*	%	*n*	%
keine Angabe	0	0	8	1	0	0	8	0
Unterkunft								
GU	165	9	188	35	28	4	325	20
dezentral	1 712	91	350	65	750	96	1 312	80
Bundesland								
Hamburg	94	5	187	35	21	3	260	16
NRW	1 223	65	168	31	589	76	802	49
RLP	174	9	113	21	87	11	200	12
Bayern	219	12	57	11	36	5	240	15
Sachsen	167	9	13	2	45	6	135	8

Beim Vergleich der Analysestichprobe 1 mit der dazugehörigen Stichprobe der ausgeschlossenen Fälle, fanden sich signifikante Abweichungen beim Standard-Item ($t(798) = 2.27$, $p = .02$, $d = 0.11$, KI für d [0.02, 0.21]) und beim Vergleich-Item ($t(1199) = 2.83$, $p = .005$, $d = 0.19$, KI für d [0.06, 0.33]) der Selbsteinschätzung der Kompetenz, die deutsche Sprache zu verstehen. Die Selbsteinschätzungen der Stichprobe der ausgeschlossenen Fälle waren auf beiden Variablen durchschnittlich etwas höher als die der Analysestichprobe. Die Effektstärken waren in beiden Fällen jedoch klein (Cohen, 1992). Weiterhin unterschieden sich die Summenwerte der Tests der kognitiven Grundfähigkeiten (DGCF-BZT, Wahrnehmungsgeschwindigkeit: $t(75) = -3.91$, $p < .001$, $d = -0.54$, KI für d [−0.83, −0.26]; DGCF-MAT, schlussfolgerndes Denken: $t(1541) = -2.34$, $p = .02$, $d = -0.39$, KI für d [−0.72, −0.06]) signifikant zwischen der Analysestichprobe und der Stichprobe der ausgeschlossenen Fälle. Die Analysestichprobe erreichte im Mittel jeweils höhere Werte als die Stichprobe der ausgeschlossenen Fälle. Dieser Vergleich ist jedoch mit Vorsicht zu betrachten, da die Fallzahlen in der Stichprobe der ausgeschlossenen Fälle mit $n = 74$ und $n = 37$ vergleichsweise klein waren und für die meisten dieser Fälle kein gültiger Summenwert für einen Deutschkompetenztest vorlag, was bedeutet, dass die Testung vorzeitig beendet wurde oder so durchgeführt wurde, dass kein Summenwert ermittelt werden konnte. Es besteht in diesen Fällen folglich die Möglichkeit, dass verschiedenartige Probleme bei der Kompetenztestung bereits bei den Deutschkompetenztests vorangehenden Tests der kognitiven Grundfähigkeiten vorhanden waren und die Ergebnisse der Tests der kognitiven Grundfähigkeiten beeinflusst und deren Validität beeinträchtigt haben.

Die Mittelwerte der Selbsteinschätzungen der Kompetenz im Verstehen der deutschen Sprache, die mittels Schieberegler-Items erfasst wurden und die Mittelwerte des Summenwerts der Can-Do-Statements zum Verstehen und Sprechen der deutschen Sprache, unterschieden sich nicht signifikant zwischen der Analysestichprobe und der Stichprobe der ausgeschlossenen Fälle.

Weiterhin gab es keine signifikanten Unterschiede zwischen den beiden Stichproben hinsichtlich Alter, Aufenthaltsdauer in Deutschland und Dauer des Schulbesuchs in Deutschland.

In der Analysestichprobe war der Anteil männlicher Teilnehmender etwas höher als in der ausgeschlossenen Stichprobe. Auffällige Unterschiede in der Stichprobenzusammensetzung fanden sich bei der Herkunft und der Muttersprache. Der Anteil Teilnehmender mit syrischer Herkunft war in der Analysestichprobe deutlich höher als in der ausgeschlossenen Stichprobe. Gleichzeitig war der Anteil von Personen mit afghanischer Herkunft in der Analysestichprobe deutlich geringer als in der ausgeschlossenen Stichprobe. Der Anteil von Personen mit irakischer Herkunft war in der Analysestichprobe nur wenig höher als in der ausgeschlossenen Stichprobe. Der Anteil von Personen mit sonstiger Herkunft war in beiden Stichproben ähnlich hoch. Ein weiterer Unterschied fand sich bei der Gruppe ohne vorliegende Angabe zur Herkunft. Deren Anteil war in der Analysestichprobe deutlich geringer als in der ausgeschlossenen Stichprobe. Ein zum Anteil der Herkunftsgruppen passendes Bild zeigte sich bei der Muttersprache. Der Anteil von Personen, die Arabisch als Muttersprache angaben, war in der Analysestichprobe deutlich höher als in der ausgeschlossenen Stichprobe. Arabisch ist in Syrien alleinige Amtssprache und im Irak neben Kurdisch Amtssprache. Der Anteil von Personen, die Persisch als Muttersprache angaben, war in der Analysestichprobe deutlich geringer als in der ausgeschlossenen Stichprobe. Persisch ist neben Paschto Amtssprache in Afghanistan. Englisch und Deutsch wurden in der Analysestichprobe häufiger als Muttersprache angegeben als in der ausgeschlossenen Stichprobe. Ein Faktor, der wahrscheinlich wesentlich zu den Unterschieden der Stichproben hinsichtlich der Herkunft und Muttersprachen beigetragen hat, ist die Reduzierung der in den Erhebungen angebotenen Befragungssprachen. Während in der Befragung der ersten Erhebungswelle acht Sprachen angeboten wurden, darunter auch Persisch, wurden die angebotenen Sprachen für die Kompetenztestung sowie die Panelbefragung auf Arabisch, Kurdisch, Englisch und Deutsch begrenzt. Personen, die ausschließlich zur Befragung ausreichende Sprachkenntnisse in Persisch, nicht jedoch in einer der vier Panelsprachen aufwiesen, konnten nicht an der Kompetenztestung teilnehmen und waren nicht Teil der Analysestichprobe. Von den Personen, deren Muttersprache nicht zu den Panelsprachen gehörte, konnten demnach nur Personen, die darüber hinaus Deutsch oder eine der anderen Panelsprachen beherrschten, in die Analysestichprobe aufgenommen werden. Demnach ist davon auszugehen, dass Teilnehmende ohne ausreichende Kenntnisse in Arabisch, Kurdisch oder Englisch abhängig von ihrer Deutschkompetenz in die Analysestichprobe aufgenommen wurden. Hinsichtlich der Deutschkompetenz der Teilnehmenden könnte die Stichprobe demzufolge selektiv sein. Ob die objektiv gemessenen Deutschkompetenzen der Teilnehmenden der Analysestichprobe im Durchschnitt etwas besser waren als die Deutschkompetenzen der von der Analysestichprobe ausgeschlossenen Fälle, ließ sich jedoch nicht prüfen, da per Stichprobendefinition für die ausgeschlossenen Fälle in der Regel keine Deutschkompetenzmaße vorlagen. Die Selbsteinschätzungen der

Deutschkompetenzen im Verstehen wichen, wie oben bereits aufgezeigt, nicht zugunsten der Analysestichprobe ab.

Deutliche Unterschiede bei der Stichprobenzusammensetzung zeigten sich darüber hinaus in Hinblick auf die Unterbringung in einer Gemeinschaftsunterkunft oder einer privaten Unterkunft, wobei Teilnehmende der Analysestichprobe häufiger privat untergebracht waren als Teilnehmende in der ausgeschlossenen Stichprobe. Dies könnte u.a. auf Umstände der Erhebung zurückzuführen sein, da insbesondere die Kompetenztestung in einem ungestörten Raum durchgeführt werden sollte. Aufgrund der vermutlich beengteren Wohnverhältnisse war diese Möglichkeit in Gemeinschaftsunterkünften wahrscheinlich seltener gegeben, sodass die Kompetenztestungen vermutlich seltener durchgeführt oder beendet wurden. Auch die Verteilung auf die fünf in die Studie eingeschlossenen Bundesländer unterschied sich deutlich zwischen der Analysestichprobe 1 und der entsprechenden Stichprobe ausgeschlossener Fälle.

Beim Vergleich der Analysestichprobe 2 mit der entsprechenden Stichprobe der ausgeschlossenen Fälle ergab sich folgendes Bild. Hinsichtlich der metrischen Variablen unterschied sich die Analysestichprobe 2 nur im Test der kognitiven Grundfähigkeiten in beiden Testteilen signifikant von der ausgeschlossenen Stichprobe (DGCF-BZT, Wahrnehmungsgeschwindigkeit: $t(1692) = -1.98$, $p = .05$, $d = -0.10$, KI für d [−0.20, −0.00]; DGCF-MAT, schlussfolgerndes Denken: $t(1541) = -4.60$, $p < .001$, $d = -0.25$, KI für d [−0.35, −0.14]). Dabei hat die Analysestichprobe im Mittel höhere Kompetenzwerte erzielt als die ausgeschlossene Stichprobe.

Die Stichprobenzusammensetzungen hinsichtlich des Geschlechts unterscheiden sich nur geringfügig. In der Analysestichprobe 2 waren die Anteile an Personen mit syrischer Herkunft und arabischer Muttersprache deutlich höher als in der ausgeschlossenen Stichprobe. Der Anteil aller weiteren Herkunftsgruppen war in der Analysestichprobe geringer als in der ausgeschlossenen Stichprobe oder in beiden Stichproben ungefähr gleich groß. Mit Ausnahme von Kurdisch war unter den Angaben zur Muttersprache der Anteil der Panelsprachen in der Analysestichprobe jeweils höher als in der ausgeschlossenen Stichprobe. Persisch wurde in der Analysestichprobe deutlich seltener als Muttersprache angegeben als in der ausgeschlossenen Stichprobe. Der Anteil an Personen, die zum Zeitpunkt der ersten Erhebungswelle in einer Gemeinschaftsunterkunft lebten, war in der Analysestichprobe deutlich geringer als in der ausgeschlossenen Stichprobe. Dies ist wie in der Analysestichprobe 1 vermutlich auf bessere Bedingungen für die Durchführung der Kompetenztestung zurückzuführen und darüber hinaus wurden Teilnehmende, die bereits in der ersten Erhebungswelle privat untergebracht waren, in der siebten Erhebungswelle vermutlich häufiger wieder erreicht, da sich die Adresse unwahrscheinlicher geändert hat. Weiterhin unterschied sich auch die Verteilung auf die fünf Bundesländer zwischen Analysestichprobe und ausgeschlossener Stichprobe deutlich.

Zusammenfassend kann nicht ausgeschlossen werden, dass in manchen Fällen die Aufnahme in die Analysestichprobe 1 von der Deutschkompetenz der Teilnehmenden abhing. Dies war der Fall, wenn Teilnehmende außer ggf. Deutsch keine Panelsprache ausreichend beherrschten, um an der Kompetenztestung und der Panelbefragung teilzunehmen. Die Selbsteinschätzungen der Deutschkompetenzen unterschieden sich jedoch nur teilweise und wenn dann mit geringer Effektstärke zugunsten der ausgeschlossenen Fälle zwischen den Stichproben. Es bleibt jedoch die Möglichkeit, dass die ausgeschlossenen Personen schlechtere Deutschkompetenzen aufwiesen und diese gleichzeitig stärker überschätzten. Beim Vergleich der Analysestichprobe 2 mit den entsprechenden ausgeschlossenen Fällen wurden keine signifikanten Unterschiede hinsichtlich der Deutschkompetenz gefunden. Für den Ausschluss aus der Analysestichprobe 2 spielte der Zufall eine deutlich größere Rolle, da die Teilnahme an der Kompetenztestung und somit die Aufnahme in die Analysestichprobe 2 in vielen Fällen davon abhing, ob die Befragung und Kompetenztestung vor dem Zeitpunkt der Entscheidung zum Umstieg auf telefonische Befragungen aufgrund der Corona-Pandemie bereits durchgeführt worden war. Für die betroffenen Fälle schien sich die Deutschkompetenz zum Zeitpunkt der ersten Erhebungswelle nicht zwischen den Stichproben zu unterscheiden. Für diejenigen Fälle, die bereits von der Analysestichprobe 1 ausgeschlossen waren, fehlten jedoch weiterhin die Ergebnisse zu den Deutschkompetenztests und konnten somit nicht im Vergleich berücksichtigt werden. Es bleibt also auch unter den von der Analysestichprobe 2 ausgeschlossenen Fällen eine Teilstichprobe, für die nicht auszuschließen ist, dass die Deutschkompetenzen geringer ausfielen als in der restlichen Stichprobe, ohne dass sich dies in den Selbsteinschätzungen manifestiert hätte.

Unterschiede zwischen den Stichproben hinsichtlich kognitiver Grundfähigkeiten konnten beim Vergleich der Analysestichprobe 1 und den entsprechenden ausgeschlossenen Fällen nicht sicher festgestellt werden, da die Fallzahlen gering waren und wie oben beschrieben davon auszugehen ist, dass bei einem relevanten Anteil der Ausschlüsse aufgrund mangelnder Deutschkompetenzwerte auch die Tests zu den kognitiven Grundfähigkeiten nicht verlässlich durchgeführt wurden. Beim Vergleich der Analysestichprobe 2 mit den entsprechenden ausgeschlossenen Fällen wurden kleine bis mittelgroße Unterschiede in den Ergebnissen der Tests zu den kognitiven Grundfähigkeiten gefunden, wobei die Analysestichprobe 2 durchschnittlich etwas besser abschnitt als die ausgeschlossenen Fälle.

Zuletzt sind die Stichproben insbesondere hinsichtlich der Art der Unterbringung sowie der Herkunft und Muttersprache der Teilnehmenden selektiv. In den Analysestichproben war der Anteil an Personen, die in einer privaten Unterkunft im Vergleich zu einer Gemeinschaftsunterkunft untergebracht waren, deutlich höher als in den Stichproben ausgeschlossener Fälle. Hinsichtlich der Herkunft war insbesondere der Anteil von Personen aus Syrien in den Analysestichproben deutlich höher als in den Stichproben ausgeschlossener Fälle und der Anteil von Personen aus

Afghanistan in den Analysestichproben deutlich geringer als in den Stichproben ausgeschlossener Fälle. Hinsichtlich der Muttersprachen war insbesondere der Anteil von arabischsprachigen Personen in den Analysestichproben deutlich höher und der Anteil von persischsprachigen Personen deutlich geringer als in den Stichproben ausgeschlossener Fälle.

Mögliche Auswirkungen der Selektivität der Stichproben auf die Ergebnisse dieser Studie und auf deren Generalisierbarkeit werden in Abschnitt 5.4.3 diskutiert.

3.2 Erhebungsablauf

Im Folgenden wird auf den Erhebungsablauf der jeweiligen Erhebungswellen und Befragungen einzeln eingegangen. Die Informationen finden sich für die erste Erhebungswelle im Methodenbericht zur ReGES-Erstbefragung C04 (Ruland et al., 2019) und für die siebte Erhebungswelle im Methodenbericht zur ReGES-Befragung C10 (Ruland et al., 2020).

3.2.1 Erhebungsablauf der ersten Erhebungswelle

Die persönlichen Befragungen der ersten Erhebungswelle fanden im Zeitraum von Ende Januar 2018 bis Ende Juni 2018 statt. Vor der Erhebung wurden die Familien schriftlich kontaktiert und erhielten Informationen über die Studie. Die Erhebungen führten Interviewerinnen und Interviewer, die neben Deutsch mindestens eine der Befragungssprachen beherrschten, persönlich bei den Familien vor Ort durch. Nach Erläuterungen zur Studie und Hinweisen auf die Freiwilligkeit und Vertraulichkeit, führten die Interviewerinnen und Interviewer mit den teilnehmenden Eltern ein computergestütztes Screening durch, um festzustellen, ob in der Familie tatsächlich ein oder mehrere Kinder lebten, die einer der Zielgruppen zugehörten. Dafür wurden insbesondere die in Abschnitt 3.1.1 genannten Teilnahmebedingungen wie z.B. das Alter, der Schulbesuch und die Panelbereitschaft nacheinander für jedes Kind geprüft sowie die Einverständnisse zur Teilnahme eingeholt. Beim Einholen der Einverständnisse wurde auch auf die für manche Jugendliche vorgesehenen „Aufgaben“ hingewiesen, ohne jedoch auf die Inhalte der Kompetenztestung einzugehen. Wurde in die Befragung der institutionellen Kontextpersonen eingewilligt, wurden im Screening auch die für die Kontaktierung der Kindertagesstätten oder Schulen relevanten Informationen zur Einrichtung und Gruppe oder Klasse abgefragt. Alle resultierenden Befragungspersonen der Familie wurden in der Kontaktverwaltung angelegt und konnten die Befragungen in weitestgehend flexibler Reihenfolge durchführen.

Die folgenden Befragungen der Eltern und Jugendlichen wurden durch die befragten Personen selbst an Tablets als Selbstausfüller beantwortet. Eine Interviewerin oder ein Interviewer war jederzeit anwesend und bereit, Rückfragen zu beantworten. Es war jederzeit möglich, am Tablet

zwischen den Befragungssprachen zu wechseln. Außerdem gab es die Möglichkeit, sich eine eingesprochene Audiodatei des jeweiligen Fragetextes und der Antwortoptionen vom Tablet vorspielen zu lassen oder dass die Interviewerin oder der Interviewer einzelne Fragen vorlas oder das gesamte Instrument als persönliches Interview durchführte. Nach Abschluss der Befragung gaben die Teilnehmenden das Tablet an den Interviewer oder die Interviewerin zurück, welcher oder welche wenige Interviewerfragen zu der vorangegangenen Befragung beantwortete, wie z.B. zu Verständnisproblemen oder zur Beteiligung Dritter an der Beantwortung der Fragen. An dieser Stelle war auch die Unterzeichnung der durch die Interviewerin oder den Interviewer zwischenzeitlich entsprechend der Angaben im Screening vorausgefüllten Einverständniserklärungen vorgesehen.

Es folgte der Übergang zur Kompetenztestung, welche an dem Laptop stattfand, an dem auch das Screening durchgeführt wurde und der über einen Touchscreen verfügte. Die Interviewerin oder der Interviewer war zunächst dazu aufgefordert, die Aufgabenbearbeitung vorzubereiten, d.h. bei Bedarf in einen ungestörten Raum mit Tisch und Sitzgelegenheiten zu wechseln und die Kopfhörer, die Tastatur und den Bildschirm des Laptops zu desinfizieren. Ansonsten wurden alle anderen noch anwesenden Personen nach Möglichkeit gebeten, den Raum zu verlassen. Zunächst wurde die oder der Jugendliche kurz über die Inhalte, die Speicherung von Logdaten und die Dauer der Aufgabenbearbeitung aufgeklärt. An dieser Stelle und unmittelbar vor dem Start der Testung bestand auch die Möglichkeit anzugeben, dass die oder der Jugendliche die Teilnahme an den Aufgaben verweigert. Weiterhin wurde überprüft, ob die richtige Instruktionssprache (Deutsch, Englisch, Arabisch oder Kurmandschi) voreingestellt war und die Spracheinstellung ggf. angepasst. Diese konnte im Verlauf der Kompetenztestung nicht mehr geändert werden. Anschließend wurde der Laptop an die Jugendliche oder den Jugendlichen übergeben, die Lautstärke der Kopfhörer eingestellt und die Kompetenztestung gestartet. Die Jugendlichen bedienten das Programm selbstständig. Die Interviewerin oder der Interviewer hielt sich im Hintergrund, sodass sie oder er bei Bedarf Fragen beantworten konnte (vgl. Obry et al., 2021).

Die Jugendlichen bearbeiteten drei verschiedene Kompetenztests in der angegebenen Reihenfolge: Einen Test der kognitiven Grundfähigkeiten (Domain General Cognitive Functions, DGCF; Lang et al., 2014) mit zwei Aufgabenteilen, einem Teil zur Wahrnehmungs- bzw. Verarbeitungsgeschwindigkeit (*Bilder-Zeichen-Test,* DGCF-BZT) und einem Teil zum schlussfolgernden Denken (*Matrizentest,* DGCF-MAT), einen Test zum rezeptiven deutschen Wortschatz (deutsche Version des Peabody Picture Vocabulary Tests, 4. Ausgabe; PPVT-4; Lenhard et al., 2015) und einen Test zum deutschen Grammatikverständnis (Test zur Überprüfung des Grammatikverständnisses; TROG-D; Fox-Boyer, 2016).

Nach Abschluss der Kompetenztests übergaben die Jugendlichen den Laptop wieder an die Interviewerin oder den Interviewer, welche wenige Interviewerfragen zur Kompetenztestung, u.a. zu Störungen und Problemen während der Testung beantworteten. Zuletzt konnten die

Teilnehmenden eine in allen acht Sprachen der ersten Befragung verfügbare Panel-App auf ihrem Smartphone installieren, über die sie nach erfolgreicher Anmeldung zu einer Online-Kurzbefragung (Welle 2) eingeladen wurden.

Zusätzlich zu den persönlichen Befragungen der Zielpersonen und Familien fanden im Rahmen der ersten Erhebungswelle auch Befragungen regionaler und institutioneller Kontextpersonen statt. Für die Analysen dieser Arbeit wurden aus diesen Befragungen jedoch ausschließlich die bei den Klassenlehrkräften in einem papierbasierten individuellen Fragebogen zur Schülerin oder zum Schüler erhobenen Mathenoten verwendet. Für nähere Informationen zur Befragung der institutionellen Kontextpersonen wird auf das Working Paper von Becker et al. (2021) verwiesen.

3.2.2 Erhebungsablauf der siebten Erhebungswelle

Die persönlichen Befragungen und Kompetenztestungen der siebten Erhebungswelle fanden im Zeitraum von Februar bis März 2020 statt. Im März 2020 mussten die persönlichen Befragungen und Kompetenztestungen bei den Familien vor Ort aufgrund der Coronapandemie vorzeitig gestoppt werden. Da die im Juni 2020 wiederaufgenommenen telefonischen Befragungen keine Kompetenztestungen umfassten, wurden die betroffenen Fälle für die Analysen dieser Arbeit ausgeschlossen und es wird an dieser Stelle ausschließlich auf den Erhebungsablauf der persönlichen Befragungen vor der pandemiebedingten Unterbrechung eingegangen. Neben einer Push-Benachrichtigung in der Panel-App für diejenigen Teilnehmenden, die die App installiert hatten, wurden die Jugendlichen im Vorfeld der Befragungen postalisch kontaktiert und erhielten verschiedene Informationen zur anstehenden Befragung sowie einen Informationsflyer mit Ergebnissen der letzten Befragungen. Vor Ort begannen die persönlichen Befragungen mit einem Kontaktmodul, das der Identifikation von Befragungspersonen im Haushalt diente und mit einer Kontaktperson der Familie durchgeführt wurde. Für die Jugendlichen startete die Befragung mit einem computergestützten persönlichen Interview (CAPI). Neben einem Intro und einem Abgleich von persönlichen Daten wie Geburtsdatum und Geschlecht wurden Fragen zum Wohnort, zu Haushaltsmitgliedern, zur Ausbildungs- und Erwerbsgeschichte, zur Teilnahme an Deutschunterricht und zum Aufenthaltsstatus gestellt. Anschließend wurde bei den Jugendlichen, die zwischenzeitlich 18 Jahre alt geworden waren, das Paneleinverständnis und das Einverständnis zur Befragung der institutionellen Kontextpersonen eingeholt. Außerdem sollte die Interviewerin oder der Interviewer Fragen u.a. zu Verständnisproblemen während der Befragung beantworten. Es folgte ein Fragebogen am Tablet, den die Jugendlichen in der Regel selbstständig ausfüllten (CASI). Wie in der ersten Befragungswelle konnten die Teilnehmenden jederzeit zwischen den angebotenen Sprachen wechseln und es bestand die Möglichkeit, dass die Interviewerin oder der Interviewer die Befragung mit der oder dem Teilnehmenden als CAPI durchführte, indem sie oder er die Fragen am Tablet vorlas

und die Antworten dokumentierte. Nach Rückgabe des Tablets an die Interviewerin oder den Interviewer war diese oder dieser dazu aufgefordert, Interviewerfragen am Laptop zu beantworten, z.B. zu Störungen und Problemen. Es folgte die Kompetenztestung, falls das Einverständnis dazu vorlag. Der Ablauf der Kompetenztestung war identisch zum Ablauf der Kompetenztestung in der ersten Erhebungswelle und ist in Abschnitt 3.2.1 bereits beschrieben. Zuletzt gab es auch zum Abschluss dieser Erhebung die Möglichkeit, gemeinsam mit der Interviewerin oder dem Interviewer die Panel-App auf dem Smartphone der Teilnehmenden zu installieren, falls noch nicht geschehen.

Auch in der siebten Erhebungswelle fanden Befragungen der institutionellen Kontextpersonen statt, welche für diese Arbeit jedoch nicht weiter von Relevanz sind.

3.3 Erhebungsinstrumente und Skalenbildung

Im Folgenden werden zu allen in den Analysen vorkommenden Variablen die zu deren Bildung verwendeten Erhebungsinstrumente dargestellt und die Kodierung der Variablen bzw. die Skalenbildung erläutert.

Fragebogenitems werden im Folgenden der Art der Präsentation am CASI-Tablet möglichst entsprechend dargestellt, sodass z.B. auch nachvollzogen werden kann, wie die Antwortoptionen angeordnet waren. Dabei werden die Items gruppiert dargestellt, wenn sie mit mehreren Items auf einer Bildschirmseite präsentiert wurden, auch wenn diese weiteren Items sonst keine Rolle in dieser Arbeit spielen. Nicht dargestellt sind hier die Buttons, die es ermöglichten, die Sprache zu wechseln, sodass das Item unmittelbar in einer anderen Sprache dargestellt wurde. Weiterhin gab es auf dem CASI-Bildschirm Buttons zum Abspielen der Audiodatei, sodass eine Audioaufnahme der entsprechenden Frage bzw. der Antwortoptionen abgespielt wurde. Unter anderem diesem Zweck dienten auch die Zahlen, mit denen die Antwortoptionen jeweils gekennzeichnet waren, sodass die Zahlen jeweils mit vorgelesen wurden und die Teilnehmenden die Antwortoption anhand der entsprechenden Zahl auswählen konnten. Die Audio-Option wurde insbesondere angeboten, damit Personen mit unzureichenden schriftsprachlichen Fähigkeiten ebenfalls an der Befragung teilnehmen konnten. Weiterhin gab es zu jeder Frage die Möglichkeit, eine der beiden Optionen *97 möchte ich nicht beantworten* oder *98 kann ich nicht beantworten* auszuwählen. Diese wurden jedoch erst eingeblendet, wenn der Weiter-Button betätigt wurde, ohne zuvor eine Antwortoption zu wählen. Die Teilnehmenden hatten außerdem die Möglichkeit, über einen Zurück-Button jederzeit zu vorherigen Fragen zurückzukehren und ihre Antworten zu ändern.

3.3.1 Items zur Selbsteinschätzung der Deutschkompetenz

In der ReGES-Studie wurden Standard-Items zur Selbsteinschätzung der Deutschkompetenz eingesetzt, wie sie z.B. auch im NEPS (s. Blossfeld & Roßbach, 2019) zum Einsatz gekommen sind, sowie drei weitere Arten von Items zur Selbsteinschätzung der Deutschkompetenz. Die Hintergründe zur Auswahl und Gestaltung der drei weiteren Arten von Selbsteinschätzungsitems wurden bereits in Kapitel 2.3 erläutert. Im Folgenden werden die Items dargestellt und es wird auf relevante Aspekte eingegangen.

3.3.1.1 Standard-Items

Die vier Standard-Items wurden im Fragebogen als Matrix dargestellt. In den Sprachen Arabisch, Persisch und Paschtu, die von rechts nach links geschrieben werden, wurde die Matrix entsprechend gespiegelt dargestellt: Der Fragetext stand jeweils rechts und die Antwortoptionen standen links davon, rechts beginnend mit *1 sehr gut* bis ganz links *5 gar nicht*. *Abbildung 2* zeigt die Standard-Items zur Selbsteinschätzung der deutschen Sprachkompetenz.

Abbildung 2. Standard-Items zur Selbsteinschätzung der deutschen Sprachkompetenz

Nun geht es darum, wie gut Sie die deutsche Sprache beherrschen.
Bitte wählen Sie in jeder Zeile eine Antwort aus.

	1 sehr gut	2 eher gut	3 eher schlecht	4 sehr schlecht	5 gar nicht
Wie gut verstehen Sie Deutsch?	O	O	O	O	O
Wie gut sprechen Sie Deutsch?	O	O	O	O	O
Wie gut lesen Sie auf Deutsch?	O	O	O	O	O
Wie gut schreiben Sie auf Deutsch?	O	O	O	O	O

Anmerkungen. ReGES Variablen t604210a – t604210d. Quelle: NEPS Variable t41030a. In der Befragung gemeinsam präsentierte, aber in den Analysen dieser Arbeit nicht verwendete Variablen, sind in grauer Schrift geschrieben.

In den Analysen dieser Arbeit habe ich hauptsächlich die erste Frage zum Verstehen der deutschen Sprache verwendet, da die Kompetenztests, die zum Vergleich als objektives Sprachkompetenzmaß herangezogen wurden, ausschließlich den rezeptiven Kompetenzbereich der gesprochenen Sprache erfassten und die mittels Selbsteinschätzung und Kompetenztests gemessenen Fähigkeiten so am besten übereinstimmten. Zumindest für einen der verwendeten Kompetenztests ist jedoch untersucht, dass er sehr hoch mit Maßen der expressiven Sprachkompetenz korreliert (Lenhard et al., 2015), weshalb in Analysen, für die die Selbsteinschätzungen latent modelliert wurden, auch das Item zum Sprechen der deutschen Sprache verwendet wurde. Für eine der Analysen

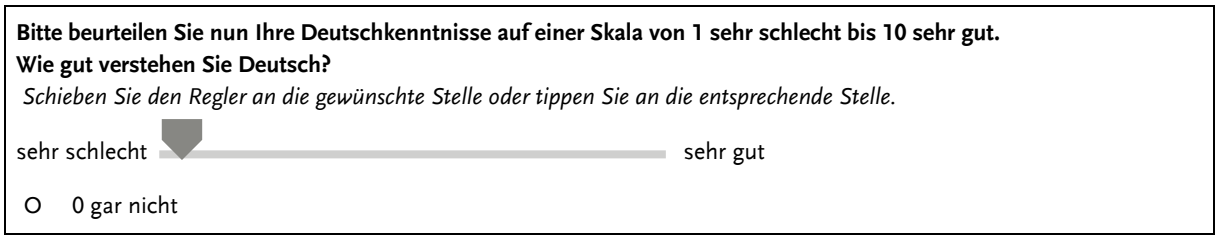
wurde der Mittelwert aus beiden Items gebildet, um mehr Kategorien zu erhalten und so der Anforderung der Maximum-Likelihood-Schätzung (ML-Schätzung) besser zu entsprechen, dass die Indikatoren der latenten Variablen eine kontinuierliche Skala aufweisen müssen (vgl. Little et al., 2022). Für eine andere Analyse wurden beide Items einzeln als Indikatoren der Standard-Selbsteinschätzung herangezogen, da mindestens zwei Indikatoren notwendig sind, um Messfehler zu schätzen und keine weiteren Indikatoren für die latente Variable zur Verfügung standen.

Die dargestellten Zahlen entsprechen nicht der Kodierung im Datensatz. Im Datensatz wurden die Antwortoptionen in umgekehrter Reihenfolge, von 1 = *gar nicht* bis 5 = *sehr gut* kodiert.

3.3.1.2 Schieberegler-Item

Der Schieberegler zur Einschätzung der Kompetenz im Verstehen der deutschen Sprache ist in *Abbildung 3* dargestellt. Für die Sprachen, die von rechts nach links geschrieben werden, war der Regler im Vergleich zur deutschen Sprache umgedreht, sodass rechts die Option *sehr schlecht* und links die Option *sehr gut* dargestellt war.

Abbildung 3. Schieberegler-Item zur Selbsteinschätzung der Kompetenz im Verstehen der deutschen Sprache



Anmerkungen. ReGES Variable t604230a.

Die vier Schieberegler-Items zum Verstehen, Sprechen, Lesen und Schreiben der deutschen Sprache haben standardmäßig nicht alle Teilnehmenden beantwortet. Die Items wurden nur einer zufällig ausgewählten Hälfte der Stichprobe vorgegeben. Die andere Hälfte der Stichprobe beantwortete stattdessen die Vergleich-Items zur Einschätzung der Deutschkompetenz. Dieses Split-Half-Design wurde gewählt, da die zur Verfügung stehende Befragungszeit begrenzt war und die Selbsteinschätzungen der Deutschkompetenzen sonst zu viel Zeit in Anspruch genommen hätten, aber gleichzeitig beide Arten von Selbsteinschätzungsitems getestet werden sollten.

Auch von den Schieberegler-Items wurde hauptsächlich das Item zum Verstehen der deutschen Sprache verwendet, mit der Ausnahme einer Analyse, für die die Selbsteinschätzung der Deutschkompetenz latent modelliert und dafür der Mittelwert aus dem Item zum Verstehen und dem Item zum Sprechen der deutschen Sprache gebildet wurde. Für das Schieberegler-Item zum Sprechen der deutschen Sprache lautete die Frage „Wie gut sprechen Sie Deutsch?“.

3.3.1.3 Vergleich-Item

Das Vergleich-Item zum Verstehen der deutschen Sprache ist in *Abbildung 4* dargestellt. Wie im vorangehenden Abschnitt bereits erläutert, wurden die Vergleich-Items zur Selbsteinschätzung der Deutschkompetenz per Zufallszuweisung nur einer Hälfte der Stichprobe vorgegeben, sodass Teilnehmende entweder die Schieberegler- oder die Vergleich-Items beantworteten.

Abbildung 4. Vergleich-Item zur Selbsteinschätzung der Kompetenz im Verstehen der deutschen Sprache

Bitte vergleichen Sie sich nun mit Gleichaltrigen, deren Muttersprache Deutsch ist.
Wie gut verstehen Sie Deutsch?

- O 1 ich verstehe gar kein Deutsch
- O 2 ich verstehe Deutsch viel schlechter als ein Deutscher
- O 3 ich verstehe Deutsch schlechter als ein Deutscher
- O 4 ich verstehe Deutsch fast genauso gut wie ein Deutscher
- O 5 ich verstehe Deutsch genauso gut wie ein Deutscher

Anmerkungen. ReGES Variable t604240a. Quelle: adaptiert nach NEPS Variable pb01030.

Genau wie bei den Schieberegler-Items wurde auch von den Vergleich-Items hauptsächlich das Item zum Verstehen der deutschen Sprache verwendet, mit der Ausnahme einer Analyse, für die die Selbsteinschätzung der Deutschkompetenz latent modelliert und dafür der Mittelwert aus dem Item zum Verstehen und dem Item zum Sprechen der deutschen Sprache gebildet wurde. Die Frage zum Vergleich-Item zum Sprechen der deutschen Sprache lautete „Wie gut sprechen Sie Deutsch?“ und in den Antwortoptionen war „ich verstehe“ jeweils durch „ich spreche“ ersetzt.

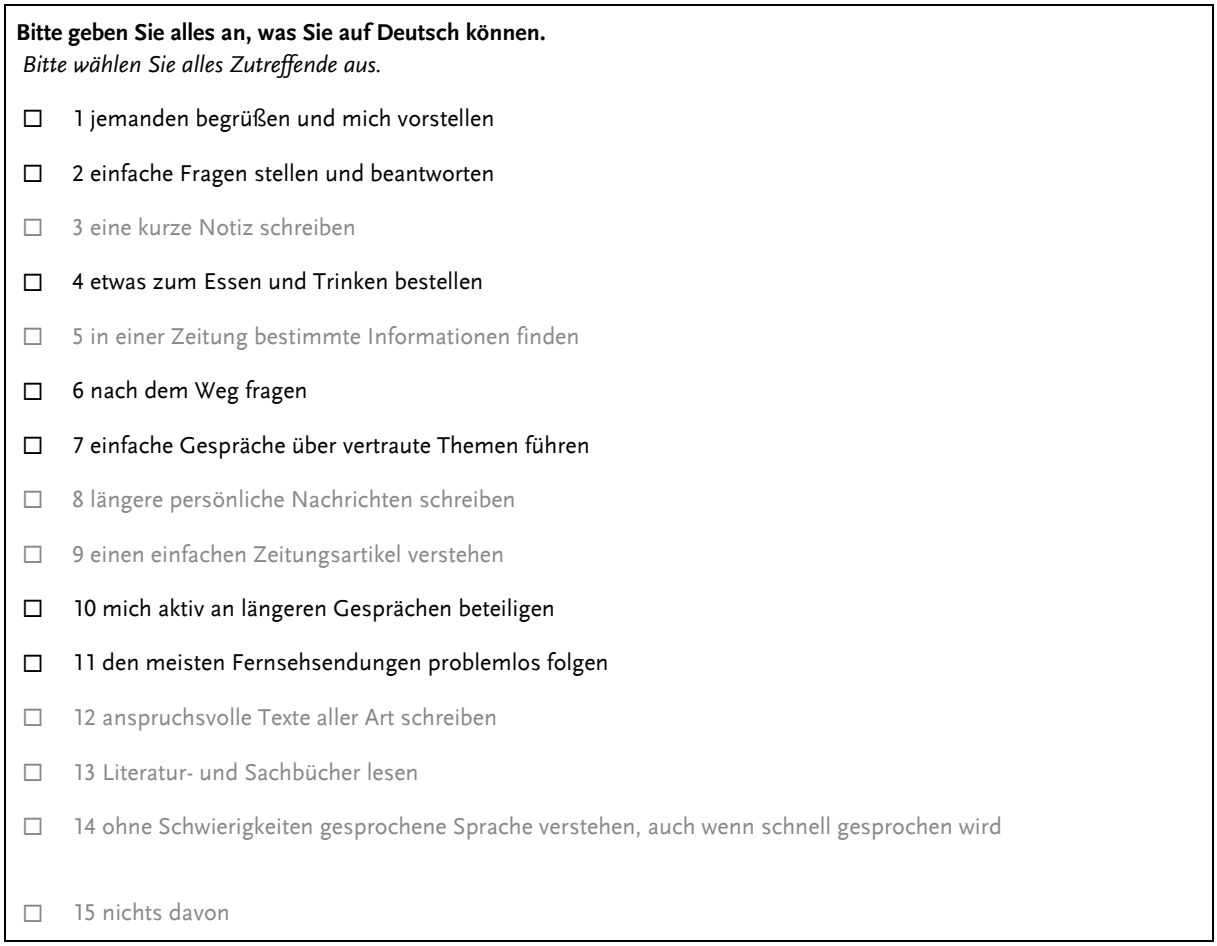
3.3.1.4 Can-Do-Statements

In *Abbildung 5* sind die Can-Do-Statements abgebildet[3]. Bei diesem Item wurden die Teilnehmenden dazu aufgefordert alle zutreffenden Statements auszuwählen, es konnten also mehrere

[3] Für die Can-Do-Statements waren in der Programmiervorlage des Fragebogens der ersten Erhebungswelle (nicht jedoch der siebten Erhebungswelle) Filterführungen enthalten, sodass zum einen Personen, die als Antwort auf die Standard-Items angegeben hatten, Deutsch nicht lesen oder nicht schreiben zu können (Antwortoption 5 gar nicht), aber zumindest eine der Fragen zum Verstehen und Sprechen mit 4 sehr schlecht bis 1 sehr gut beantwortet haben, eine gekürzte Version der Can-Do-Statements vorgegeben bekamen. Diese gekürzte Version enthielt nur die Statements zum Verstehen und Sprechen der deutschen Sprache. Diese Konstellation ist jedoch nur in zwei Fällen der Analysestichprobe 1 eingetreten. Zum anderen wurden Personen, die als Antwort auf die Standard-Items angegeben hatten, Deutsch weder verstehen noch sprechen zu können (Antwortoption 5 gar nicht), die Can-Do-Statements in der Befragung nicht vorgegeben. Diese Antwortkombination ist jedoch nur in einem Fall der Analysestichprobe 1 aufgetreten, für den der Summenwert zu den Can-Do-Statements als fehlender Wert kodiert wurde.

Antwortoptionen gewählt werden. Im Datensatz wurde jedes Statement einzeln in einer Variable erfasst und mit 1 kodiert, falls es gewählt wurde und mit 0 kodiert, falls es nicht gewählt wurde.

Abbildung 5. Can-Do-Statements zur Selbsteinschätzung der Deutschkompetenz


Bitte geben Sie alles an, was Sie auf Deutsch können.
Bitte wählen Sie alles Zutreffende aus.

☐ 1 jemanden begrüßen und mich vorstellen
☐ 2 einfache Fragen stellen und beantworten
☐ 3 eine kurze Notiz schreiben
☐ 4 etwas zum Essen und Trinken bestellen
☐ 5 in einer Zeitung bestimmte Informationen finden
☐ 6 nach dem Weg fragen
☐ 7 einfache Gespräche über vertraute Themen führen
☐ 8 längere persönliche Nachrichten schreiben
☐ 9 einen einfachen Zeitungsartikel verstehen
☐ 10 mich aktiv an längeren Gesprächen beteiligen
☐ 11 den meisten Fernsehsendungen problemlos folgen
☐ 12 anspruchsvolle Texte aller Art schreiben
☐ 13 Literatur- und Sachbücher lesen
☐ 14 ohne Schwierigkeiten gesprochene Sprache verstehen, auch wenn schnell gesprochen wird

☐ 15 nichts davon

Anmerkungen. ReGES Variablen t604250a – t604250n. Quelle: adaptiert nach Deutscher Volkshochschul-Verband e.V. (2011). Die Statements, die nicht zur Summenwertbildung verwendet wurden, sind in grauer Schrift geschrieben.

Da in dieser Arbeit hauptsächlich die Kompetenzen im Verstehen der deutschen Sprache betrachtet werden, wurden für die Analysen nur die Statements ausgewählt, die sich auf Kompetenzen im Verstehen beziehen, wobei die meisten der Statements zum Verstehen auch Kompetenzen im Sprechen der deutschen Sprache umfassen und diese somit nicht getrennt betrachtet werden können. Aufgrund der oben bereits berichteten hohen Korrelation zwischen rezeptiven und expressiven Sprachkompetenzen sollte dies jedoch unproblematisch sein. Bei den Statements zum Verstehen und Sprechen der deutschen Sprache handelt es sich um die Statements, die in *Abbildung 5* mit den Ziffern 1, 2, 4, 6, 7, 11 und 14 gekennzeichnet sind. Aus den Angaben zu den einzelnen Statements zum Verstehen und Sprechen wurde ein Summenwert gebildet, der die Anzahl der ausgewählten Statements wiedergibt. Bevor der Summenwert gebildet werden konnte, musste die Skala jedoch hinsichtlich ihrer Qualität geprüft werden, woraufhin das 14. Statement von der

Berechnung des Summenwerts ausgeschlossen wurde. In die Summenwertbildung gingen also nur die Statements mit den Ziffern 1, 2, 4, 6, 7 und 11 ein. Das Vorgehen und die Ergebnisse der Skalenqualitätsprüfung werden in Abschnitt 4.1.4 berichtet.

3.3.2 Deutschkompetenztests

In der ReGES-Studie wurden Kompetenzen der jugendlichen Flüchtlinge im Verständnis der deutschen Sprache anhand zweier etablierter Kompetenztests erfasst. In den folgenden Abschnitten beschreibe ich die Tests und Besonderheiten bei der Anwendung in der ReGES-Studie. Ich gehe darauf ein, welche Kompetenzwerte ich für die Analysen eingesetzt habe und gebe einen Überblick über fehlende Kompetenzwerte.

3.3.2.1 Rezeptiver Wortschatz

Einige der folgenden sowie weitere Informationen zur Messung des rezeptiven Wortschatzes in der ersten Erhebungswelle der ReGES-Studie haben meine Kollegin, meine Kollegen und ich bereits in einem Working Paper (Obry et al., 2021) veröffentlicht. Für diese Arbeit wesentliche Aspekte fasse ich im Folgenden zusammen.

Den rezeptiven Wortschatz haben wir in der ReGES-Studie mit einer adaptierten deutschen Variante (Lenhard et al., 2015) der vierten Ausgabe des *Peabody Picture Vocabulary Tests* (PPVT-4; Dunn & Dunn, 2007) erhoben. Beim PPVT-4 werden der Testperson vier farbige Bilder und ein gesprochenes deutsches Wort präsentiert. Die Person ist dazu angeleitet, das zu dem gesprochenen Wort passende Bild auszuwählen. Der getestete Wortschatz ist vielfältig, da die Testitems einer großen Bandbreite verschiedener Inhaltskategorien zuzuordnen sind. Insgesamt besteht der PPVT-4 aus 228 Items, die in 19 Itemsets á 12 Items mit aufsteigender Schwierigkeit angeordnet sind. Der Test beginnt mit einer Übungsphase, gefolgt von einer Testphase. Die Testphase ist mit einem Abbruchkriterium versehen, sodass nur die Itemsets vorgegeben werden, die für die Testperson weder zu schwierig, noch zu einfach sind (Lenhard et al., 2015).

In der ReGES-Studie haben wir eine technologiebasierte Variante des Tests verwendet, die bis auf bestimmte Ausnahmen entsprechend der Vorgaben im Testmanual programmiert ist. Die Instruktionen und die Stimuli wurden den Teilnehmenden in Form von Audioaufnahmen über Kopfhörer präsentiert. Die Instruktionen wurden in der Sprache präsentiert, die die Teilnehmenden zu Beginn der Testung ausgewählt hatten. Das Itemset, mit dem die Jugendlichen im Anschluss an eine erfolgreich bearbeitete Übungsphase starteten, entsprach nicht dem Set, das im Testmanual für die Altersgruppe empfohlen wurde. Für die jugendlichen Flüchtlinge haben wir nach Rücksprache mit einem der Testautoren, Prof. Dr. Wolfgang Lenhard, das Set 5 als Startset gewählt, da die Startsets im Testmanual anhand einer Stichprobe festgelegt wurden, die überwiegend aus

deutschen Muttersprachlerinnen und Muttersprachlern bestand. Deshalb befürchteten wir, dass das übliche Startset für einen großen Teil der Stichprobe sehr schwierig und demotivierend sein könnte. Darüber hinaus haben wir aufgrund der Unangemessenheit für die Zielstichprobe und der Gefahr von Retraumatisierungen, das Zielwort „Detonation“ durch das Zielwort „Eruption“ ersetzt und bei weiteren drei Items wurde ein Distraktorbild durch ein anderes Bild ersetzt (vgl. Obry et al., 2021).

Für Analysen mit Daten ausschließlich der ersten Erhebungswelle habe ich die im Rahmen der Analysen zum Working Paper (Obry et al., 2021) geschätzten und auch im Scientific-Use-File enthaltenen Weighted Likelihood Estimates (WLEs) verwendet. Die WLEs lagen für all diejenigen Fälle mit vollständiger Testbearbeitung vor, für die auch ein PPVT-4-Summenwert entsprechend den Vorgaben im Manual berechnet werden konnte. Zusätzlich zu den Fällen mit vollständiger Testbearbeitung wurden WLEs für diejenigen Jugendlichen ($n = 58$) geschätzt, die trotz Abbruch der Testung fünf oder mehr Itemsets komplett bearbeitet haben, weil wir davon ausgingen, dass für diese Fälle genügend Informationen vorlagen, um Personenwerte anhand eines Item Response Theorie-Modells zu schätzen. Weil für diese Fälle WLEs, jedoch keine Summenwerte berechnet werden konnten, und somit für mehr Fälle WLEs als Summenwerte vorlagen, habe ich die WLEs gegenüber den Summenwerten bevorzugt.

Für die Schätzung der WLEs haben wir folgende Entscheidungen getroffen. Wie es auch für die Berechnung des Summenwerte vorgesehen ist (Lenhard et al., 2015), wurden alle Items, die unterhalb des *Bodensets* lagen, also nicht vorgegeben wurden, weil sie als zu leicht für die Testperson galten, als richtig gelöst gewertet. Alle Items oberhalb des *Deckensets*, die nicht vorgegeben wurden, weil sie als zu schwer für die Testperson galten, wurden als falsch gelöst gewertet. In den Fällen, die die Testung nach der Bearbeitung von mindestens fünf Sets abgebrochen haben, wurden alle schwierigeren nicht erreichten Items für die Schätzung der WLEs ignoriert, da das Abbruchkriterium noch nicht erreicht wurde und deshalb keine Annahme dazu getroffen werden konnte, ob die Items richtig oder falsch beantwortet worden wären. Weiterhin wurden für die Schätzung der WLEs nur die ersten 16 Itemsets des PPVT-4 berücksichtigt, jedoch nicht die letzten drei, da diese aufgrund des Abbruchkriteriums von sehr wenigen (max. 100) Personen erreicht wurden und somit nur wenige gültige Werte vorlagen (s. Obry et al., 2021).

Für Analysen mit Daten der siebten Erhebungswelle habe ich ausschließlich entsprechend dem Testmanual berechnete Summenwerte verwendet. Da im Rasch-Modell „all das, was in den Daten über die Fähigkeit einer Person steckt, durch die Anzahl der gelösten Aufgaben erschöpfend repräsentiert [wird]“ (Eid & Schmidt, 2014, S. 162), hätte es für die Analysen keinen Mehrwert gehabt, auf dem Rasch-Modell basierende WLEs analog zur ersten Erhebungswelle zu schätzen, sofern dadurch keine zusätzlichen Fälle hätten gewonnen werden können. Da von den Fällen ohne gültigen PPVT-Summenwert in der siebten Erhebungswelle keiner mindestens fünf Sets

vollständig bearbeitet hat, hätten auf dem oben beschriebenen Weg keine zusätzlichen Fälle durch die Schätzung von WLEs gewonnen werden können.

Für die Analysen zur Untersuchung der Hypothese 5b, bei der die Veränderungen der Kompetenzen zwischen den beiden Messzeitpunkten betrachtet wurden, musste ich für beide Erhebungswellen auf die Summenwerte zum PPVT-4 zurückgreifen, damit die Skalierung der Werte zwischen beiden Wellen vergleichbar war.

Verschiedene Gründe, wie z.B. Testabbrüche oder technische Probleme können dazu geführt haben, dass kein gültiger WLE oder Summenwert für einen Fall vorlag. *Tabelle 3* bietet eine Übersicht über die Gründe für den Ausschluss von Fällen und die jeweilige Anzahl der betroffenen Fälle, zum einen für die jeweilige gesamte Stichprobe von 2 016 jugendlichen Flüchtlingen in der ersten Erhebungswelle und 794 jugendlichen Flüchtlingen in der siebten Erhebungswelle, für die die Kompetenztestung administriert wurde und zum anderen für die beiden Analysestichproben mit 1 877 und 778 Fällen. Dabei ist zu beachten, dass Fälle u.a. dann von den Analysen dieser Arbeit ausgeschlossen wurden, wenn weder ein gültiger Kompetenzwert zum PPVT-4 noch ein gültiger Kompetenzwert zum TROG-D vorlag. Die Tabelle habe ich aus unserem Working Paper (Obry et al., 2021) übernommen und angepasst und erweitert.

Tabelle 3. Gründe für fehlende PPVT-4 Kompetenzwerte sowie Anzahl der Fälle und Anteile in Prozent

Grund für fehlenden WLE bzw. Summenwert	Erhebungswelle 1		Erhebungswelle 7	
	administrierte Testungen (*N* = 2 016)	Analysestichprobe 1 (*N* = 1 877)	administrierte Testungen (*N* = 794)	Analysestichprobe 2 (*N* = 778)
–91 Kompetenztestung abgebrochen[a]	< 5 Sets bearbeitet: 511 (25%) >= 5 Sets bearbeitet[f] 58 (3%)	< 5 Sets bearbeitet: 399 (21%) >= 5 Sets bearbeitet[f] 58 (3%)	6 (1%)	2 (< 1%)
–90 nicht spezifizierbar fehlend[b]	19 (1%)	8 (< 1%)	0 (0%)	0 (0%)
–27 keine valide Aussage möglich: Abbruchkriterium in Set 1 erreicht[c]	3 (< 1%)	3 (< 1%)	2 (< 1%)	1 (< 1%)
–24 Übungsphase nicht bestanden[d]	20 (1%)	7 (< 1%)	13 (2%)	2 (< 1%)
–21 technische Probleme/Absturz[e]	12 (1%)	11 (1%)	0 (0%)	0 (0%)

Grund für fehlenden WLE bzw. Summenwert	Erhebungswelle 1		Erhebungswelle 7	
	administrierte Testungen (*N* = 2 016)	Analysestichprobe 1 (*N* = 1 877)	administrierte Testungen (*N* = 794)	Analysestichprobe 2 (*N* = 778)
Gesamte Ausschlüsse WLE	565 (28%)	428 (23%)	-[g]	-[g]
Gesamte Ausschlüsse Summenwert	623 (31%)	486 (26%)	21 (3%)	5 (1%)

Anmerkungen. WLE = Weighted Likelihood Estimate. Adaptiert nach Obry et al. (2021).
[a] Kein Wert vorliegend, da die Testanwendung über das Testleitermenü abgebrochen wurde oder Testteile über das Testleitermenü übersprungen wurden. [b] Kein Wert vorliegend aus unbekanntem Grund; es konnte bei diesen Fällen nicht rekonstruiert werden, ob es sich um Testabbrüche oder Systemabstürze handelte. [c] Das Abbruchkriterium (>= 8 Fehler) wurde bereits in Set 1 erreicht. Entsprechend der Vorgabe im Testmanual wird die Person als nicht testbar mit dem PPVT betrachtet und es wird kein Summenwert/WLE angegeben. [d] Keine Testdurchführung, da die Übungsphase nicht bestanden wurde. [e] Kein Wert vorliegend, da die Testanwendung abgestürzt ist. [f] Für Fälle, die den PPVT-4 abgebrochen aber mindestens 5 Sets vollständig bearbeitet haben, wurden WLEs für die Erhebungswelle 1 geschätzt aber keine Summenwerte berechnet. [g] Für die Erhebungswelle 7 wurden keine WLEs geschätzt.

Der Anteil an Fällen, die die Testung in der ersten Erhebungswelle abgebrochen haben, war sehr hoch, weshalb ich an dieser Stelle mögliche Gründe dafür zusammenfasse. Da die Kompetenztestung erst im Anschluss an die in der ersten Erhebungswelle sehr langen Befragungen durchgeführt wurde, ist es möglich, dass die zeitliche Belastung zu diesem Zeitpunkt schon sehr hoch war und dies zu vermehrten Testabbrüchen geführt hat. Zudem mussten die Jugendlichen im DGCF-MAT bei jedem der drei Aufgabenblöcke die vorgegebenen 3 Minuten Bearbeitungszeit abwarten, bevor sie zum nächsten Teil übergehen konnten. Dies könnte zu Irritationen und einem Abbruch einzelner Testteile oder des gesamten Testmoduls geführt haben, in das der PPVT-4 (nicht jedoch der TROG-D) ebenfalls integriert war. Von den Fällen, für die kein PPVT-4-Summenwert vorlag, weil die Kompetenztestung abgebrochen wurde, lag auch in 385 Fällen der Stichprobe aller administrierten Testungen und in 294 Fällen der Analysestichprobe 1 kein Summenwert zum DGCF-MAT aufgrund eines Testabbruchs vor. In der siebten Erhebungswelle gingen dem PPVT-4 jedoch dieselben Tests voraus und dies scheint dort nicht vermehrt zu Testabbrüchen geführt zu haben. Weiterhin könnte die Testsituation in der ersten Erhebungswelle für einen Teil der Stichprobe ungewohnt gewesen sein.

3.3.2.2 Grammatikverständnis

Das Grammatikverständnis wurde in der ReGES-Studie mit dem *Test zur Überprüfung des Grammatikverständnisses* (TROG-D; Fox-Boyer, 2016) gemessen, welcher auf dem englischsprachigen *Test for Reception of Grammar* (TROG; Bishop, 1989) basiert. Den Testpersonen wird ein Satz

auditiv präsentiert und diese sind wie beim PPVT-4 dazu aufgefordert, das dazu passende Bild aus vier Bildern auszuwählen. Die Distraktoren sind im Vergleich zu dem Inhalt des dargebotenen Satzes grammatisch oder lexikalisch nur leicht verändert. Auf diese Weise wird das Verständnis grammatischer Strukturen wie z.B. Plural oder Passiv getestet. Das Vokabular der Sätze ist einfach. Im originalen Test wird jede der 21 grammatischen Strukturen in einem Block mit vier Items geprüft. Die Blöcke sind in aufsteigender Schwierigkeit angeordnet und es gibt ein Abbruchkriterium (Fox-Boyer, 2016).

In der ReGES-Studie haben wir hingegen eine auf 48 Items gekürzte Version des TROG-D ohne das in der originalen Version vorgesehene Abbruchkriterium für die Kohorte der jugendlichen Flüchtlinge eingesetzt. Die Itemauswahl entspricht der gekürzten Version des TROG-D beim Einsatz zum ersten Messzeitpunkt der Startkohorte 2 im NEPS (Lorenz et al., 2017). Die ersten drei Blöcke sind in dieser Version vollständig enthalten, alle folgenden Blöcke wurden auf zwei der Items reduziert. Obwohl im originalen Test vorgesehen ist, dass Kinder ab 7 Jahren mit dem vierten Block starten, bearbeiteten die Jugendlichen den Test vollständig und starteten mit dem ersten Block. Wie auch beim PPVT-4 haben wir eine digitale Version des Tests eingesetzt, die bis auf die genannten Abweichungen entsprechend den Vorgaben im Testmanual programmiert ist. Auch hier wurden die Instruktionen in der ausgewählten Sprache und die Items auf Deutsch als Audioaufnahmen über Kopfhörer präsentiert.

Für die Analysen habe ich für beide Erhebungswellen den auch im Scientific-Use-File enthaltenen Summenwert verwendet, der die Summe der richtig gelösten Items für alle Fälle, die den Test vollständig bearbeitet haben, wiedergibt und aus diesem das Item Nr. 4 in Block O des TROG-D herausgerechnet. Dieses Item wurde nach sorgfältiger Itemanalyse ausgeschlossen, was in Abschnitt 4.1.2 genauer erläutert wird. Der Summenwert wurde anders als im Testmanual vorgesehen berechnet, wonach die Anzahl korrekter Blöcke zu zählen wäre. Auch hier hätte es für die Analysen keinen Mehrwert gehabt, auf dem Rasch-Modell basierende WLEs analog zum PPVT-4 zu schätzen, sofern dadurch nicht beispielsweise zusätzliche Fälle gewonnen worden wären. Da der TROG-D seltener vorzeitig abgebrochen wurde als der PPVT-4 und insgesamt kürzer ist, hätten auf diesem Wege kaum oder keine zusätzlichen Fälle ähnlich zum Vorgehen beim PPVT-4 für die Analysen gewonnen werden können.

Für den Fall, dass der Test nicht vollständig bearbeitete wurde, konnte kein Summenwert berechnet werden. Der Summenwert fehlt somit für diejenigen Fälle, die den Test vorzeitig abgebrochen haben oder zwei Blöcke in Folge keine Antwort auf alle Items gegeben haben. In diesem Fall wurde der Test, wie im Manual vorgegeben, automatisch vorzeitig abgebrochen. Einen Überblick über die Anzahl fehlender Werte aus den beschriebenen Gründen gibt *Tabelle 4*.

Tabelle 4. Gründe für fehlende TROG-D Summenwerte sowie Anzahl der Fälle und Anteile in Prozent

Grund für fehlenden Summenwert	Erhebungswelle 1		Erhebungswelle 7	
	administrierte Testungen (*n* = 2 016)	Analysestichprobe 1 (*N* = 1 877)	administrierte Testungen (*n* = 794)	Analysestichprobe 2 (*N* = 778)
–91 Kompetenztestung abgebrochen[a]	186 (9%)	67 (4%)	7 (1%)	3 (< 1%)
–90 nicht spezifizierbar fehlend[b]	1 (< 1%)	0 (0%)	27 (3%)	26 (3%)
–20 keine valide Aussage möglich: automatischer Abbruch wegen Non-Response[c]	24 (1%)	9 (< 1%)	16 (2%)	5 (1%)
Gesamte Ausschlüsse	211 (10%)	76 (4%)	50 (6%)	34 (4%)

Anmerkungen. [a] Kein Wert vorliegend, da die Testanwendung über das Testleitermenü abgebrochen wurde. [b] Kein Wert vorliegend aus unbekanntem Grund; es konnte bei diesen Fällen nicht rekonstruiert werden, ob es sich um Testabbrüche oder Systemabstürze handelte. [c] Der Test wurde automatisch abgebrochen, da in zwei aufeinanderfolgenden Blöcken keine Antworten auf die Items gegeben wurden.

3.3.2.3 Kombinierter Deutschkompetenzscore

Für die Berechnung verschiedener Analysen wurden die beiden Kompetenzwerte des PPVT-4 und des TROG-D zu einem kombinierten Kompetenzwert für die Kompetenz im Hörverstehen der deutschen Sprache zusammengefasst. Dazu wurden beide Kompetenzwerte z-skaliert und anschließend der Mittelwert aus beiden z-skalierten Werten berechnet. In der ersten Erhebungswelle wurden die WLEs des PPVT-4 für die Berechnung des kombinierten Deutschkompetenzscores herangezogen, ansonsten wurden jeweils die Summenwerte verwendet. Damit der kombinierte Deutschkompetenzscore auch für die Berechnung der Differenz zwischen Selbsteinschätzung und objektivem Kompetenzmaß geeignet war, wurde er wie im Folgenden erläutert weiter transformiert.

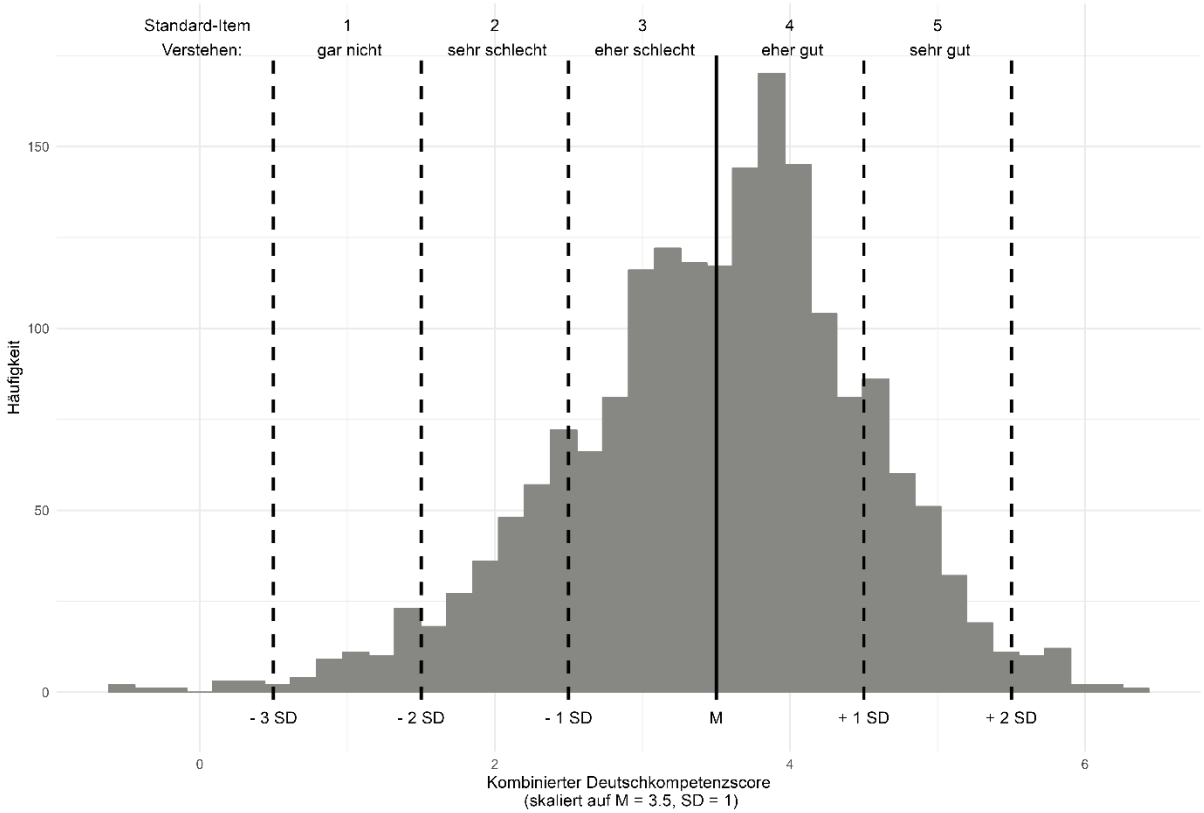
Um für Hypothese 1b die Differenz zwischen Selbsteinschätzung und objektivem Kompetenzmaß berechnen zu können, müssen diese auf derselben Skala liegen. Selbsteinschätzung und objektives Kompetenzmaß liegen bereits auf derselben Skala, wenn Teilnehmende schätzen, wie viele Punkte oder welchen Prozentrang innerhalb einer bestimmten Gruppe sie in einer Prüfung oder einem Test erreicht haben (z.B. Kruger & Dunning, 1999). Hier schätzten die Teilnehmenden ihre Deutschkompetenz jedoch auf einer Likert-Skala ein, welche nicht auf derselben Skala wie der aus

den z-Werten gemittelte Deutschkompetenzscore liegt. In diesem Fall muss zunächst definiert werden, bei welchem objektiven Kompetenzwert welche Selbsteinschätzung die richtige wäre. Die Entscheidung, welchem Kompetenzwert welche Selbsteinschätzung zugeordnet wird, wirkt sich darauf aus, welche Selbsteinschätzungen als Über- oder Unterschätzung gewertet werden und auf das Ausmaß der angenommenen Über- oder Unterschätzung und sollte entsprechend überlegt getroffen werden. Entsprechend dieser Zuordnung müssen die Skalen anschließend transformiert werden (vgl. Artelt & Rausch, 2014).

Grundlage für die Zuordnung von Selbsteinschätzungen zu Kompetenzwerten sind verschiedene Annahmen. Hier wurde angenommen, dass die Gesamtstichprobe der jugendlichen Flüchtlinge als Referenzgruppe für die Selbsteinschätzungen diente. Alternativ wären verschiedene Referenzgruppen denkbar gewesen, wie z.B. eine gleichaltrige Stichprobe von Muttersprachlerinnen und Muttersprachlern. Andere jugendliche Flüchtlinge als Referenzgruppe anzunehmen ist angemessen, da sich Personen auf subjektiven Likert-Skalen relativ zu den Standards einschätzen, die sie für sich bzw. die Gruppe, der sie angehören, annehmen (Biernat, 2005; Biernat & Manis, 1994; Heine et al., 2002). Während der Befragung der ReGES-Studie war die Zugehörigkeit zur Gruppe der Flüchtlinge salient und wurde vermutlich überwiegend als Vergleichsstandard angenommen. Die Gesamtstichprobe als Referenzrahmen ergibt sich ergänzend aus Sicht der Forschungsfrage, ob die Selbsteinschätzungen als Kompetenzmaß innerhalb der Stichprobe differenzieren. Für die Annahme der Gruppe der jugendlichen Flüchtlinge als Referenz sprach also einerseits die implizite Interpretation der Selbsteinschätzungsskala und andererseits der Fokus der Forschungsfrage auf die Differenzierbarkeit der Kompetenzmaße innerhalb der untersuchten Gruppe.

Weiterhin wurde die Annahme getroffen, dass sich die positiven Selbsteinschätzungskategorien *4 eher gut* und *5 sehr gut* auf überdurchschnittliche Kompetenzwerte im Vergleich zur Gesamtstichprobe bezogen und die negativen Selbsteinschätzungskategorien *1 gar nicht*, *2 sehr schlecht* und *3 eher schlecht* auf die unterdurchschnittlichen Kompetenzwerte. Dabei wurde angenommen, dass der Mittelwert der Gesamtstichprobe in der Mitte zwischen den Kategorien *3 eher schlecht* und *4 eher gut* lag und dass der Abstand zwischen zwei benachbarten Selbsteinschätzungskategorien einer Standardabweichung des kombinierten Deutschkompetenzscores entsprach. Die beschriebene Zuordnung ist in *Abbildung 6* grafisch dargestellt. Beispielsweise wurde demnach von einer oder einem Teilnehmenden, die oder der in den Kompetenztests so abgeschnitten hat, dass ihr oder sein kombinierter Deutschkompetenzscore eine halbe Standardabweichung über dem Mittelwert der Gesamtstichprobe lag, erwartet, dass sie oder er ihre oder seine Kompetenz, Deutsch zu verstehen als *4 eher gut* einstufte.

Abbildung 6. Veranschaulichung der Definition einer genauen Selbsteinschätzung und der Umskalierung des kombinierten Deutschkompetenzscores



Anmerkungen. Die in grau dargestellte Verteilung stellt die Häufigkeitsverteilung des kombinierten Deutschkompetenzscores in der ersten Erhebungswelle der ReGES-Studie dar. Der Mittelwert sowie die Abstände von 1 bis 3 Standardabweichungen zum Mittelwert sind als senkrechte durchgezogene bzw. gestrichelte Linie eingezeichnet. Oben ist die Zuordnung der Kategorien des Standard-Items zur Selbsteinschätzung der Kompetenz im Verstehen der deutschen Sprache zu den Wertebereichen des kombinierten Deutschkompetenzscores dargestellt. $N = 1\,877$.

Um den Annahmen zur Zuordnung von Selbsteinschätzungen zu Kompetenzwerten zu entsprechen, wurden die Werte des kombinierten Deutschkompetenzscores durch eine lineare Transformation an die Skala der Selbsteinschätzung angepasst, sodass der Mittelwert des kombinierten Deutschkompetenzscores in der Stichprobe bei $M = 3.5$, also dem oben definierten durchschnittlichen Wert der Selbsteinschätzungsskala, und die Standardabweichung weiterhin bei $SD = 1.0$ lag, also dem oben definierten Abstand zwischen zwei benachbarten Selbsteinschätzungskategorien. Das heißt, wenn wie im Beispiel oben der kombinierte Deutschkompetenzscores eine halbe Standardabweichung über dem Mittelwert der Gesamtstichprobe lag und die Selbsteinschätzungskategorie *4 eher gut* gewählt wurde, lag sowohl der Wert für den kombinierten Deutschkompetenzscore als auch der Wert für die Selbsteinschätzung bei 4.0.

Mithilfe des so skalierten kombinierten Deutschkompetenzscores konnte eine Variable für die Verzerrung als die Differenz zwischen dem Wert des kombinierten Deutschkompetenzscores und

der Selbsteinschätzung berechnet werden. Im Beispiel oben erhielt die Verzerrungs-Variable den Wert 4.0 − 4.0 = 0.0. Es lag eine genaue Selbsteinschätzung vor. Wenn als weiteres Beispiel jemand zur Einschätzung der eigenen Deutschkompetenz die Kategorie *eher gut* (4.0) wählte und einen kombinierten Deutschkompetenzscore von 3.0 erreichte, erhielt die Variable den Wert 4.0 − 3.0 = 1.0. Es lag eine Überschätzung vor. Bei negativer Differenz wurde von einer Unterschätzung ausgegangen.

3.3.3 Leistung in Mathematik

Zur Erfassung der Leistung in Mathematik wurden die selbstberichteten Schulnoten aus dem letzten Zeugnis verwendet, sowie die Lehrerangabe zur Mathematiknote im letzten Halbjahreszeugnis. Eine Studie zur Akkuratheit von selbstberichteten Zeugnisnoten an 866 Schülerinnen und Schülern an unterschiedlichen deutschen Schulformen der 7. und 8. Klassenstufe ergab, dass die selbstberichteten Noten der letzten Klassenarbeit in Mathematik und der letzten Zeugnisnote in Mathematik sehr hoch mit den entsprechenden Lehrerangaben korrelierten ($r = .90$ und $r = .88$). Im Durchschnitt überschätzten die Schülerinnen und Schüler ihre Noten leicht, jedoch mit sehr geringer Effektstärke. Auch im Vergleich mit einem Leistungstest waren die Angaben der Schülerinnen und Schüler nicht weniger valide als die der Lehrerinnen und Lehrer (Dickhäuser & Plenter, 2005). Da sich die hier betrachtete Stichprobe jedoch von der Stichprobe in der Studie von Dickhäuser und Plenter (2005) unterscheidet, kann die hohe Genauigkeit der selbstberichteten Mathematiknoten hier nicht einfach angenommen werden. Die geringere Vertrautheit der jugendlichen Flüchtlinge mit dem deutschen Notensystem könnte z.B. die Erinnerung an die letzten Zeugnisnoten beeinträchtigen. Deshalb scheint es sinnvoll, die Lehrerangaben zusätzlich heranzuziehen, auch wenn diese nur in wenigen Fällen vorliegen.

In *Abbildung 7* ist die Frage zur Schulnote im letzten Zeugnis in Mathematik dargestellt.

Abbildung 7. Fragebogenitem zur Mathematiknote im letzten Zeugnis

Welche Noten hatten Sie im letzten Zeugnis im Fach Mathematik?

O 1 sehr gut

O 2 gut

O 3 befriedigend

O 4 ausreichend

O 5 mangelhaft

O 6 ungenügend

O 7 trifft nicht zu: Ich habe noch kein Zeugnis bekommen

Anmerkungen. ReGES Variable t3324010. Quelle: adaptiert nach NEPS Variable p724102.

In *Abbildung 8* ist die Frage zu den Noten der Schülerin oder des Schülers aus dem letzten Halbjahreszeugnis in verschiedenen Fächern aus dem schriftlichen Fragebogen zur Schülerin oder zum Schüler abgebildet, den die Klassenlehrkräfte beantwortet haben. Ganzzahlige Noten wurden für die Analysen übernommen, Angaben mit + und –, wie z.B. 2+ oder 2– wurden umkodiert indem 0.25 addiert oder subtrahiert wurde, z.B. zu 1.75 bzw. 2.25. Weitere uneindeutige Angaben wurden als fehlende Werte kodiert.

Beide Variablen zu den Mathematiknoten wurden mit −1 multipliziert, sodass höhere Werte mit einer höheren Leistung in Mathematik einhergehen, was die Interpretation der Ergebnisse erleichtert.

Abbildung 8. Fragebogenitem zu Noten u.a. in Mathematik aus dem letzten Halbjahreszeugnis aus dem schriftlichen Fragebogen für die Klassenlehrkräfte zur Schülerin oder zum Schüler

Bitte tragen Sie im Folgenden die Noten der Schülerin oder des Schülers aus dem letzten Halbjahreszeugnis ein.
Bitte jede Zeile ausfüllen.
Wenn Notenpunkte (z.B. 0–15) erteilt werden, bitte in einstellige Noten (1–6) umrechnen.

	Note	Noch keine Noten in diesem Fach erhalten	Wird noch nicht in diesem Fach unterrichtet
a) Deutsch	☐	☐	☐
b) Mathematik	☐	☐	☐
c) Englisch	☐	☐	☐

Anmerkungen. ReGES Variable e3324020.

3.3.4 Engagement beim Deutschlernen

In der ReGES-Studie wurde die Nutzung verschiedener Möglichkeiten zum Deutschlernen erfasst. Diese Informationen habe ich in dieser Arbeit als Indikator des Engagements beim Deutschlernen herangezogen. Die Frage dazu, welche Möglichkeiten die Teilnehmenden außerhalb von institutionellen Angeboten genutzt haben, um ihre Deutschkompetenzen zu verbessern, ist in *Abbildung 9* dargestellt. Bei dieser Frage konnten eine oder mehrere der Antwortoptionen 1 bis 8 oder die Antwortoption 9 *nichts davon* gewählt werden. Aus den einzelnen Angaben sollte ein Summenwert gebildet werden, der die Anzahl der ausgewählten Statements wiedergibt. Bevor der Summenwert gebildet werden konnte, musste die Skala jedoch hinsichtlich ihrer Qualität geprüft und ggf. Items ausgeschlossen werden. Die Ergebnisse der Skalenqualitätsprüfung sowie Angaben zur Berechnung des Summenwerts werden in Abschnitt 4.1.5 berichtet.

Abbildung 9. Fragebogenitem zur Nutzung von Möglichkeiten zum Deutschlernen

Haben Sie eine oder mehrere der folgenden Möglichkeiten genutzt, um Ihre Deutschkenntnisse zu verbessern?
Bitte wählen Sie alles Zutreffende aus.

- ☐ 1 Unterricht durch Eltern, Verwandte, Freunde oder Bekannte
- ☐ 2 Sprachlern-CD, Sprachlern-App, Sprachkurs im Internet oder Ähnliches
- ☐ 3 Internet in deutscher Sprache
- ☐ 4 Sprachlehrbücher
- ☐ 5 Fernsehen auf Deutsch
- ☐ 6 deutsche Zeitungen und Bücher lesen
- ☐ 7 Gespräche und Kontakt mit Personen, die Deutsch sprechen
- ☐ 8 andere Möglichkeiten

- ☐ 9 nichts davon

Anmerkungen. ReGES Variablen t624231a – t624231z. Quelle: adaptiert nach TNS Infratest Sozialforschung (2016) Frage 130.

3.3.5 Schlussfolgerndes Denken

Kognitive Fähigkeiten zum schlussfolgernden Denken wurden in der ReGES-Studie mit dem DGCF-MAT (Lang et al., 2014) erfasst, welcher einen traditionellen Matrizen-Test mit variierender Schwierigkeit darstellt. Der Test umfasst drei Aufgabenblöcke mit jeweils vier Matrizenaufgaben. Für jeden Aufgabenblock hatten die Teilnehmenden 3 Minuten Zeit, um diesen zu lösen. Für jeden Aufgabenblock wurde ein Summenwert der richtigen Antworten gebildet. In die Analysen gingen die drei Summenwerte der Aufgabenblöcke ein.

3.3.6 Teilnahme an einem Deutschkurs

Zu einer Variablen zur Deutschkursteilnahme wurden die Fragen zur aktuellen Teilnahme an einem Deutschkurs (s. *Abbildung 10*) und zur Teilnahme an einem Deutschkurs in der Vergangenheit (s. *Abbildung 11*) zusammengefasst. Die Variable wurde mit 1 = *ja* kodiert, wenn eine der beiden Fragen mit ja beantwortet wurde und mit 0 = *nein*, wenn beide Fragen mit nein beantwortet wurden. Die Frage zur Teilnahme an einem Deutschkurs in der Vergangenheit wurde im Fragebogen nur gestellt, wenn die Frage zur aktuellen Teilnahme mit *4 nein* beantwortet wurde.

Abbildung 10. Fragebogenitem zur aktuellen Teilnahme an einem Deutschkurs

Erhalten Sie aktuell Unterricht speziell für Flüchtlinge oder Migrantinnen und Migranten zum Deutschlernen?

O 1 ja, in der Schule

O 2 ja, außerhalb der Schule

O 3 ja, in der Schule und außerhalb der Schule

O 4 nein

Anmerkungen. ReGES Variable t6242200.

Abbildung 11. Fragebogenitem zur vergangenen Teilnahme an einem Deutschkurs

Haben Sie in der Vergangenheit schon mal solchen Deutschunterricht für Flüchtlinge oder Migrantinnen und Migranten erhalten?

O 1 ja

O 2 nein

Anmerkungen. ReGES Variable t6242201.

3.3.7 Teilnahme an einem Deutschtest

In *Abbildung 12* ist die Frage zur Teilnahme an einem Deutschtest dargestellt. Die Antwortoptionen wurden für die Analysen mit 1 = *ja* und 0 = *nein* kodiert.

Abbildung 12. Fragebogenitem zur Teilnahme an einem Deutschtest

Haben Sie schon mal einen Deutschtest gemacht, bei dem Ihnen ein Deutschniveau von A1 bis C2 bescheinigt wurde?

O 1 ja

O 2 nein

Anmerkungen. ReGES Variable t6242120. Quelle: adaptiert nach TNS Infratest Sozialforschung (2016).

4 Analysen und Ergebnisse

Das Kapitel zu den Analysen und Ergebnissen gliedert sich in sechs Teile. In Kapitel 4.1 werden Analysen und Ergebnisse zur Qualität der verwendeten Skalen berichtet. Dazu gehören sowohl die etablierten Kompetenztests, die hier bei einer besonderen Stichprobe eingesetzt wurden, als auch die Can-Do-Statements und die Items zur Nutzung von Möglichkeiten zum Deutschlernen, die in der Form noch nicht näher untersucht wurden. In Kapitel 4.2 werden univariate deskriptive Statistiken zu den später verwendeten Variablen berichtet. In Kapitel 4.3 folgen die Ergebnisse zu bivariaten Zusammenhängen zwischen den verwendeten Variablen. Ab dem Kapitel 4.4 werden die Analysen und Ergebnisse der Hypothesenprüfungen erläutert, beginnend mit den Ergebnissen zu den Hypothesen 1a – 1c zur Genauigkeit der Selbsteinschätzungen, wobei zuerst auf die Diskrimination, anschließend auf die allgemeine Verzerrung und abschließend auf die Variation der Selbsteinschätzungen eingegangen wird. In Kapitel 4.5 werden die Analysen und Ergebnisse zu den Hypothesen 2a – 2e zu den Einflussfaktoren von Selbsteinschätzungen beschrieben. In Kapitel 4.6 wird auf die Analysen und Ergebnisse zu den verschiedenen Arten von Selbsteinschätzungsitems eingegangen. Zuerst geht es um die Hypothesen 3a und 3b zu den Schieberegler-Items, anschließend um die Hypothesen 4a und 4b zu den Vergleich-Items und abschließend um die Hypothesen 5a und 5b zu den Can-Do-Statements.

4.1 Skalenanalysen

In die Analysen dieser Arbeit wurden Ergebnisse von Kompetenztestungen zum rezeptiven deutschen Wortschatz, zum Grammatikverständnis und zum schlussfolgernden Denken einbezogen. Bei allen drei eingesetzten Verfahren handelt es sich um etablierte und ausführlich hinsichtlich der Testgüte geprüfte Verfahren (PPVT-4: s. Lenhard et al. (2015); TROG-D: s. Fox-Boyer (2016); DGCF-MAT: s. Lang et al. (2014)). Jedoch wurden sie hier bei einer speziellen Stichprobe eingesetzt, die in die Evaluierung der Testgüte nicht einbezogen war und teilweise wurden kleine Änderungen an den Tests vorgenommen, die in Kapitel 3.3 beschrieben sind. Deshalb wird hier vorab über Ergebnisse zur Testgüte der eingesetzten Verfahren in beiden Erhebungswellen der ReGES-Studie berichtet. Die Testgüte des PPVT-4 in der ersten Erhebungswelle haben eine Kollegin, zwei Kollegen und ich ausführlich evaluiert und die Ergebnisse in einem Working Paper veröffentlicht (Obry et al., 2021). Diese sind in Abschnitt 4.1.1 zusammengefasst. Anhand der Ergebnisse aller Kompetenztests in den beiden Erhebungswellen, für die noch keine ausführlichen Skalenanalysen vorlagen, habe ich kürzere Skalenanalysen durchgeführt und mich auf die interne

Konsistenz, die Itemschwierigkeiten und die Trennschärfen der Items konzentriert. Die Ergebnisse werden in den Abschnitten 4.1.1, 4.1.2 und 4.1.3 berichtet.

Weiterhin wurden in dieser Arbeit die Can-Do-Statements zum Verstehen und Sprechen der deutschen Sprache und die Items zur Nutzung von Möglichkeiten beim Deutschlernen eingesetzt, welche jeweils eine Skala zur Messung eines Konstrukts bilden sollen, jedoch bisher nicht hinsichtlich ihrer Skalenqualität evaluiert wurden. Auch diese Skalen wurden vor ihrer weiteren Verwendung hinsichtlich der internen Konsistenz sowie der Itemschwierigkeiten und Trennschärfen der Items geprüft. Um sicherzustellen, dass die Items dieser noch nicht etablierten Skalen zu einem Summenwert zusammengefasst werden können, wurde darüber hinaus die Eindimensionalität der Skalen geprüft. Anhand der Ergebnisse der Skalenanalysen habe ich ggf. Items für die Bildung eines Summenwerts ausgeschlossen. Die Ergebnisse der Skalenanalysen der Can-Do-Statements und der Nutzung von Möglichkeiten zum Deutschlernen werden in den Abschnitten 4.1.4 und 4.1.5 berichtet.

Die internen Konsistenzen, Itemschwierigkeiten und Trennschärfen der jeweiligen Skalen wurden mithilfe der alpha-Funktion des psych-Pakets Version 2.2.5 (Revelle, 2022) in R Version 4.0.3 (R Core Team, 2020) analysiert. Zur Prüfung der Eindimensionalität der Skalen habe ich die mirt-Funktion des mirt-Paktes Version 1.37.1 (Chalmers, 2012) in R Version 4.0.3 (R Core Team, 2020) verwendet. Mithilfe dieser Funktion konnten ein- und mehrfaktorielle Modelle der explorativen Faktorenanalyse an die dichotomen Itemdaten angepasst und anschließend die Modelle miteinander verglichen werden (s. Finch & French, 2015).

4.1.1 PPVT-4

Zur Evaluierung der Skalenqualität des PPVT-4 in der ersten Erhebungswelle haben wir im Rahmen des o.g. Working Papers (Obry et al., 2021) verschiedene psychometrische Eigenschaften des Tests betrachtet und uns an dem empfohlenen Vorgehen von Pohl und Carstensen (2012) sowie der Skalenqualitätsprüfung einer anderen Version des PPVT im NEPS (Fischer & Durda, 2020) orientiert. Wir evaluierten Itemschwierigkeiten, die Gegenüberstellung von Itemschwierigkeiten und Personenfähigkeiten, Item-Fit, Trennschärfen, Reliabilität, Differential-Item-Functioning (DIF), Rasch-Homogenität und Eindimensionalität. Dazu wurden nur die ersten 16 Itemsets berücksichtigt, da die letzten drei Itemsets nicht mehr als 100 gültige Antworten enthielten.

Wie auch für die Berechnung des Summenwerts wurden für die Skalenanalysen die Items unterhalb des Bodensets bzw. oberhalb des Deckensets als richtig bzw. falsch beantwortet kodiert. Da die Items den Teilnehmenden jedoch tatsächlich nicht vorgegeben wurden, kann dies mit Verzerrungen einhergehen, worauf im Folgenden ggf. hingewiesen wird. Die Reliabilität des PPVT-4 wird aus diesem Grund hier zusätzlich so berechnet, dass nur Daten in die Analysen eingehen, die

auf tatsächlichen Antworten der Teilnehmenden beruhen (vgl. Lenhard et al., 2015). Inwiefern das Ausgehen von der Annahme, dass die nicht administrierten Items unterhalb des Bodensets gelöst und die nicht administrierten Items oberhalb des Deckensets nicht gelöst worden wären, sich auch auf die Qualität der Summenwerte und WLEs in dieser Stichprobe ausgewirkt haben könnte, was jedoch nicht einfach prüfbar ist, wird in Abschnitt 5.4.2 diskutiert.

Die Items deckten einen breiten Schwierigkeitsbereich ab. Der Anteil korrekter Antworten variierte zwischen 2% und 99% (M = 47%, SD = 34%) und auch die mittels Rasch-Modell geschätzten Itemschwierigkeiten deuteten auf eine breite Verteilung hin (Obry et al., 2021). Die extremen Werte können jedoch teilweise dadurch beeinflusst sein, dass z.B. die Items in den höchsten Sets von sehr wenigen Teilnehmenden bearbeitet wurden und für alle anderen Teilnehmenden als nicht gelöst kodiert wurden. Die Breite der Verteilung der Itemschwierigkeiten wäre also wahrscheinlich etwas geringer, wenn die Teilnehmenden alle Items bearbeitet hätten. Auch bei einer etwas geringeren Breite der Verteilung der Itemschwierigkeiten ist jedoch davon auszugehen, dass für alle Teilnehmenden Items mit angemessener Schwierigkeit zur Verfügung standen und der Test im Kompetenzbereich der Teilnehmenden gut differenzieren konnte.

Der Item-Fit wurde anhand von Weighted-Mean-Square-Statistiken (WMNSQ) evaluiert. Dieser Index beschreibt die Abweichung der beobachteten Wahrscheinlichkeit einer richtigen Antwort von der durch das Modell implizierten Wahrscheinlichkeit einer richtigen Antwort bei einem gegebenen Fähigkeitsniveau (Pohl & Carstensen, 2012). Einen als stark zu bewertenden Item-Misfit mit WMNSQ-Werten über 1.2 wiesen 27 Items auf, was ca. 14% der evaluierten Items entspricht (Obry et al., 2021). WMNSQ-Werte über 1.2 sind als Verletzung der Modellanpassung zu werten und deuten auf eine schlechte Differenzierung der jeweiligen Items zwischen Personen hin (vgl. Pohl & Carstensen, 2012). Die Itemcharakteristikkurven der betroffenen Items wiesen jedoch auf eine ausreichende Passung zum Rasch-Modell hin (Obry et al., 2021). Da die *beobachtete* Wahrscheinlichkeit einer richtigen Antwort durch die Kodierung als richtig bzw. falsch gelöst ohne tatsächliche Vorgabe des Items nicht dem wahren Wert entspricht, der beobachtet worden wäre, wenn das Item von allen Teilnehmenden bearbeitet worden wäre, können auch die Item-Fit-Statistiken ungenau sein. Zum einen hat dies womöglich die Schätzung der Fähigkeitsparameter verzerrt, was zum Misfit mancher Items beigetragen haben könnte (Obry et al., 2021). Zum anderen wurden insbesondere in den höheren Itemsets, die von wenigeren Teilnehmenden bearbeitet und somit in vielen Fällen einheitlich als falsch gelöst kodiert wurden, keine Items mit starkem Item-Misfit beobachtet. Hätten die Teilnehmenden die Items tatsächlich bearbeitet, wäre die WMNSQ-Statistik womöglich auch für manche dieser Items auffällig gewesen. Die Trennschärfen der Items lagen im Durchschnitt bei r_{it} = .44 (SD = .15), wobei 16 Items (ca. 8%) eine kleinere Korrelation als r_{it} = .20 zwischen dem Item und dem um das Item korrigierten Testscore aufwiesen. Nach der von Pohl und Carstensen (2012) formulierten Daumenregel sind Korrelationen < .20 als problematisch

zu werten, Korrelationen > .20 als akzeptabel und Korrelationen > .30 als gut. Das heißt, dass 92% der Items eine akzeptable oder gute Trennschärfe aufwiesen. Auch die Trennschärfen könnten jedoch durch die Kodierung nicht vorgegebener Items als richtig oder falsch gelöst verzerrt sein.

Für die Reliabilität wurden im Rahmen des Working Papers folgende Werte geschätzt: EAP-Reliabilität = .98 und WLE-Reliabilität = .98. Demnach war die Reliabilität des Tests sehr gut, wurde jedoch wahrscheinlich überschätzt, da alle nicht vorgegebenen Items oberhalb des Deckensets als falsch gelöst gewertet wurden (Obry et al., 2021).

Um ein Maß für die Reliabilität zu erhalten, das weniger durch die Kodierung der nicht administrierten Items verzerrt ist, habe ich ergänzend zum Working Paper den Cronbachs α-Koeffizienten für beide Erhebungswellen auf die annähernd gleiche Art und Weise ermittelt, wie es im Testmanual des PPVT-4 (Lenhard et al., 2015) beschrieben ist. Das heißt, ich habe Cronbachs α jeweils für Intervalle des Tests, z.B. Set 1 bis Set 6, ermittelt und dafür nur Fälle einbezogen, die dieses Testintervall komplett bearbeitet haben und alle Fälle ausgeschlossen, die einen Teil des Intervalls wegen des Abbruchkriteriums nicht bearbeitet haben. Die Größe der Testintervalle, für die Cronbachs α ermittelt werden sollte, wurde anhand folgender Überlegungen bestimmt. Es sollten möglichst wenige Intervalle gebildet werden, die Intervalle sollten jedoch so klein sein, dass die Anzahl der Teilnehmenden, die alle Itemsets der Intervalle bearbeitet haben, eine ausreichend große Stichprobe bildeten, um die interne Konsistenz sinnvoll bestimmen zu können. Dafür wurde von einem Richtwert von mindestens $n = 100$ Fällen pro Intervall ausgegangen. Teilnehmende, die weniger Itemsets als die definierte Intervallgröße bearbeitet haben, mussten von den Analysen ausgeschlossen werden, weil sie für kein Itemintervall alle Items bearbeitet haben, während Teilnehmende, die mehr Itemsets als die definierte Intervallgröße bearbeitet haben, in die Analyse der internen Konsistenz mehrerer Intervalle einbezogen wurden. Im Durchschnitt haben Teilnehmende in der ersten Erhebungswelle $M = 8.50$ Itemsets ($SD = 2.76$) bearbeitet. Ausgehend von einer Normalverteilung hat somit ca. die Hälfte der Teilnehmenden 8.50 oder mehr Itemsets bearbeitet. Würde man eine Intervallgröße von 8 oder 9 Itemsets wählen, würde somit ca. die Hälfte der Stichprobe von den Analysen ausgeschlossen werden. Ebenfalls ausgehend von einer Normalverteilung, haben ca. 84% der Stichprobe mindestens $M - 1 \cdot SD$ Itemsets bearbeitet, was im konkreten Fall $8.50 - 2.76 = 5.74$ also ca. 6 Itemsets entspricht. Dies scheint ein guter Kompromiss zu sein, bei dem ein großer Teil der Stichprobe in die Analysen der internen Konsistenz einbezogen wird und die Itemintervalle noch eine relevante Anzahl an Sets umfassen. Bei einer Intervallgröße von 6 Itemsets ergeben sich für die erste Erhebungswelle Stichprobengrößen im Bereich von $n = 118$ bis $n = 831$ für die jeweiligen Intervalle. Um die Vergleichbarkeit für die siebte Erhebungswelle zu gewährleisten, wurden auch hier Intervalle von 6 Itemsets gewählt, auch wenn die Teilnehmenden in der siebten Erhebungswelle durchschnittlich ca. 1 Itemset mehr bearbeitet haben. Für die siebte Erhebungswelle ergeben sich so Stichprobengrößen von $n = 82$ bis $n = 595$. Einen Überblick über

Cronbachs α für beide Erhebungswellen und alle möglichen Intervalle von 6 Itemsets mit $n > 100$ gibt *Tabelle 5*. Die jeweils berücksichtigte Stichprobe umfasste ausschließlich Fälle, die alle einbezogenen Itemsets vollständig bearbeitet haben. Somit waren die einbezogenen Fälle für die Intervalle auch vom Leistungsniveau der Teilnehmenden abhängig. Insgesamt war die interne Konsistenz der Skala in beiden Erhebungswellen akzeptabel bis sehr gut, fiel jedoch niedriger aus als in der Normstichprobe, wo der Wert für Cronbachs α für Intervalle von 9 Itemsets zwischen α = .82 und α = .95 lag (Lenhard et al., 2015).

Tabelle 5. Interne Konsistenz verschiedener Itemintervalle des PPVT-4

Itemintervall	Erhebungswelle 1		Erhebungswelle 7	
	n	Cronbachs α	*n*	Cronbachs α
1–6	242	.81	-	-
2–7	299	.75	105	.78
3–8	754	.72	355	.77
4–9	715	.78	382	.78
5–10	831	.85	595	.88
6–11	589	.81	477	.84
7–12	474	.80	401	.82
8–13	393	.85	350	.86
9–14	255	.78	242	.76
10–15	198	.79	195	.78
11–16	118	.77	136	.68

Anmerkungen. Cronbachs α wurde für die angegebenen Itemintervalle jeweils anhand der Daten aller Fälle ermittelt, die das Itemintervall vollständig bearbeitet haben. Für Itemintervalle, die von weniger als 100 Fällen vollständig bearbeitet wurden, wurde die interne Konsistenz nicht ermittelt.

Mittels DIF wurde die Messinvarianz zwischen verschiedenen Gruppen untersucht, also ob sich die Lösungswahrscheinlichkeit bestimmter Items bei gleicher Fähigkeit zwischen den Gruppen unterschied und betroffene Items somit eine der Gruppen bevorteilten (vgl. Pohl & Carstensen, 2012). Es wurden jeweils alle Items identifiziert, deren Wert für die absolute Differenz zwischen den geschätzten Schwierigkeiten 0.4 Logit überstieg und für die somit von mindestens beachtlichem DIF zwischen den betrachteten Gruppen ausgegangen wurde (vgl. Pohl & Carstensen, 2012). Für das Geschlecht wurden 35 Items mit einem beachtlichen DIF-Wert von mindestens 0.4 Logit identifiziert, was 18% der berücksichtigten 192 Items entspricht. Im Vergleich mit dem Haupteffektmodell bevorzugte das AIC das DIF-Modell, während das BIC das Haupteffektmodell

bevorzugte. Im Durchschnitt erzielten die männlichen Jugendlichen bessere Testergebnisse als die weiblichen Jugendlichen (Haupteffekt = -0.33 Logit; Obry et al., 2021).

Für das Herkunftsland wurde zwischen Jugendlichen syrischer und nicht-syrischer Herkunft unterschieden. In diesem Fall wurden 36 Items mit beachtlichen DIF-Werten von mindestens 0.4 Logit gefunden. Dies entspricht einem Anteil von ca. 19% der Items. Das Haupteffektmodell wurde gegenüber dem DIF-Modell durch beide Gütekriterien bevorzugt. Der durchschnittliche Fähigkeitsunterschied zwischen beiden Gruppen war gering (Haupteffekt = 0.05 Logit; Obry et al., 2021).

Zuletzt wurde DIF hinsichtlich dem Bildungshintergrund der Eltern untersucht, wobei drei Gruppen unterschieden wurden: Eltern mit mindestens einem Sekundarabschluss, Eltern mit höchstens einem Primarabschluss und Fälle, für die es keine Angaben zum Bildungsabschluss der Eltern gab. Beim Vergleich der Jugendlichen aus den Haushalten mit höherer Bildung mit denen aus Haushalten mit niedrigerer Bildung gab es 67 Items mit beachtlichem DIF, was ca. 35% der Items entspricht. Beim Vergleich der Jugendlichen aus Haushalten mit höherer Bildung mit denjenigen ohne Angaben gab es 31 Items mit beachtlichem DIF, was ca. 16% der Items entspricht und im Vergleich der Jugendlichen aus Haushalten mit niedrigerer Bildung mit denjenigen aus Haushalten ohne Angaben gab es 57 Items mit beachtlichem DIF, was ca. 30% der Items entspricht. AIC und BIC bevorzugen im Vergleich jedoch das Haupteffektmodell gegenüber dem DIF-Modell. Bezüglich der Haupteffekte hat sich gezeigt, dass Jugendliche aus gebildeteren Elternhäusern bessere Leistungen erzielten als Jugendliche aus weniger gebildeten Elternhäusern (Haupteffekt = –0.81 Logit) und als Jugendliche aus Haushalten ohne Angaben zur Bildung (Haupteffekt = –0.46 Logit). Jugendliche aus Haushalten mit niedrigerer Bildung erzielten im Durchschnitt geringere Leistungen als Jugendliche aus Haushalten ohne Angaben zur Bildung (Haupteffekt = 0.34 Logit; Obry et al., 2021).

Insgesamt scheint ein großer Anteil der Items beachtliche DIF-Werte für mindestens einen der Gruppenvergleiche aufzuweisen. Es bestünde die Möglichkeit, nach zusätzlicher inhaltlicher Betrachtung einzelne Items mit sehr hohen DIF-Werten für die Berechnung der Testwerte auszuschließen. Insgesamt kann nach den Modellvergleichen jedoch in allen Fällen das Haupteffektmodell gegenüber einem Modell, das DIF zulässt, bevorzugt werden. Da in dieser Arbeit darüber hinaus keine Vergleiche zwischen den hier betrachteten Gruppen durchgeführt werden, für die entsprechender DIF problematisch wäre, soll stattdessen die Vergleichbarkeit des Summenscores mit dem der Normstichprobe beibehalten werden und dafür alle Items in die Analysen miteinbezogen werden.

Die Berechnung des Summenscores impliziert zudem, dass das Rasch-Modell gilt, nach dem sich die Items nur anhand ihres Schwierigkeitsparameters unterscheiden und die Diskriminationsparameter aller Items den gleichen Wert haben. Um die Rasch-Homogenität, also die Annahme gleicher Diskriminationsparameter, zu prüfen, wurde zusätzlich ein zweiparametrisches

Testmodell geschätzt, bei dem auch die Diskriminationsparameter zwischen den Items variieren konnten, und mit dem Rasch-Modell verglichen (vgl. Fischer & Durda, 2020). Auf Grundlage der Informationskriterien AIC und BIC müsste das zweiparametrische Modell gegenüber dem einparametrischen Rasch-Modell bevorzugt werden. Jedoch entspricht das einparametrische Modell dem Konstruktionsrational des Tests, weshalb entschieden wurde, das Rasch-Modell trotzdem weiter anzuwenden. Auch hier ist anzunehmen, dass die Diskriminationsparameter durch die Kodierung nicht bearbeiteter Items als richtig oder falsch gelöst beeinflusst wurden (Obry et al., 2021).

Eine auf den Korrelationen der Residuen des Rasch-Modells basierende Prüfung der Dimensionalität bestätigte die Eindimensionalität des Tests (vgl. Gnambs, 2017).

Da für den PPVT-4 in der siebten Erhebungswelle noch keine ausführliche Skalenqualitätsprüfung durchgeführt wurde, habe ich neben der oben berichteten internen Konsistenz die Trennschärfen und die Schwierigkeiten der Items untersucht. Um die Trennschärfen zu ermitteln und die Itemschwierigkeiten anhand derselben Stichprobe berechnen zu können, wurden auch hierfür die nicht vorgegebenen Items oberhalb des Deckensets als falsch und die nicht vorgegebenen Items unterhalb des Bodensets als richtig gelöst kodiert. Itemsets, die von nicht mehr als 100 Teilnehmenden bearbeitet wurden, wurden von den Analysen ausgeschlossen. Dies betraf alle Items ab Set 17. Die mittlere Trennschärfe lag bei $r_{it} = .43$ ($SD = .17$, $Min = .02$, $Max = .80$). Die Trennschärfen der meisten Items lagen somit in einem akzeptablen oder guten Bereich. Die mittlere Itemschwierigkeit, also der mittlere Anteil korrekter Antworten, lag bei $M = .54$ ($SD = .34$, $Min = .03$, $Max = 1.00$), ist also angemessen für die Teilnehmenden und so breit verteilt, dass sie den Kompetenzbereich der Teilnehmenden gut abdeckt. Eine Übersicht über die einzelnen Trennschärfen und Itemschwierigkeiten ist in *Tabelle A 1* in Anhang A dargestellt. Da das Item Nummer 9 im zweiten Set keine Varianz aufwies, wurde es von den Analysen zur Trennschärfe und den Itemschwierigkeiten ausgeschlossen.

Zusammenfassend sind die interne Konsistenz, die Trennschärfen und die Schwierigkeitsverteilung der Items des PPVT-4 zufriedenstellend, auch die Eindimensionalität des Tests wurde bestätigt. Manche Items weisen bei Gruppenvergleichen DIF auf, wobei der Tests insgesamt relativ faire Vergleiche zwischen den untersuchten Gruppen zuzulassen scheint. Problematisch erscheint, dass manche Items einen Misfit aufweisen. Zudem scheint die Annahme gleicher Diskriminationsparameter der Items nicht erfüllt zu sein. Trotz dieser Problematiken wurde entschieden, für die weiteren Analysen den entsprechend den Vorgaben im Testmanual berechneten Summenwert zu verwenden und auch die WLEs zu nutzen, für deren Schätzung die nicht bearbeiteten Items entsprechend dem Konstruktionsrational des Tests kodiert wurden (vgl. Obry et al., 2021).

4.1.2 TROG-D

Zur Prüfung der Testgüte des TROG-D wurde Cronbachs α als Maß für die interne Konsistenz bzw. Reliabilität des Tests sowie die Schwierigkeit und Trennschärfe der Items ermittelt. In der ersten Erhebungswelle betrug die interne Konsistenz für die 1 801 Fälle mit gültigem TROG-D-Summenwert α = .85. Die mittlere Trennschärfe lag bei r_{it} = .29 (SD = .15, Min = –.11, Max = .51), wobei das Item Nr. 4 in Block O der TROG-D eine negative Trennschärfe aufwies. Die Trennschärfen der meisten Items lagen jedoch in einem akzeptablen bis guten Bereich. Die mittlere Itemschwierigkeit lag bei M = .71 (SD = .30, Min = .07, Max = 1.00). Die Items waren also tendenziell einfach für die Stichprobe, deckten jedoch einen breiten Schwierigkeitsbereich ab. In der siebten Erhebungswelle betrug die interne Konsistenz für die 744 Fälle mit gültigem TROG-D-Summenwert α = .86. Die mittlere Trennschärfe lag bei r_{it} = .34 (SD = .11, Min = .01, Max = .51) und somit ebenfalls für die meisten Items in einem akzeptablen bis guten Bereich. Das Item Nummer 4 in Block O des TROG-D, das in der ersten Erhebungswelle eine negative Trennschärfe hatte, hatte in der siebten Erhebungswelle eine ebenfalls problematische Trennschärfe von r_{it} = .01. Die mittlere Itemschwierigkeit lag bei M = .76 (SD = .29, Min = .08, Max = 1.00), d.h. der Test fiel den Teilnehmenden ein wenig leichter als in der ersten Erhebungswelle.

Aufgrund der schlechten Trennschärfe-Werte wurde das Item Nummer 4 in Block O des TROG-D auch inhaltlich betrachtet. Es beschreibt eine Situation, die im Alltag sehr unüblich ist. Möglicherweise wurde es deshalb auch von Personen mit besseren Kompetenzen häufig nicht verstanden und differenziert folglich schlecht zwischen den Kompetenzen der Teilnehmenden. Da die Jugendlichen nur 48 Items der insgesamt 84 Items des TROG-D bearbeitet haben und der Summenwert anders berechnet wurde, als es im Manual des Tests vorgegeben ist, wurde in diesem Fall der Ausschluss dieses Items mit sehr problematischen Itemkennwerten bevorzugt gegenüber der Beibehaltung des originalen Tests. Aus diesem Grund wurde das Item Nummer 4 in Block O von den Analysen dieser Arbeit ausgeschlossen und aus dem im Scientific-Use-File enthaltenen Summenwert des TROG-D herausgerechnet.

Eine erneute Prüfung der Güte des um das eine Item reduzierten Tests ergab Folgendes: In der ersten Erhebungswelle wurde für die interne Konsistenz ein Wert von Cronbachs α = .85 berechnet. Die Trennschärfe lag im Mittel bei r_{it} = .30 (SD = .14, Min = .04, Max = .51) und somit für die meisten Items in einem akzeptablen bis guten Bereich. Wenige Items hatten eine sehr niedrige Trennschärfe, dies betraf jedoch hauptsächlich sehr schwierige Items, welche beibehalten wurden, da Items in extremen Schwierigkeitsbereichen häufig eine niedrige Trennschärfe haben und die schwierigen Items zur Differenzierung im oberen Kompetenzbereich benötigt wurden. Außerdem war die Trennschärfe dieser Items in der siebten Erhebungswelle höher als in der ersten Erhebungswelle und größtenteils akzeptabel. Die Schwierigkeit lag im Mittel bei M = .72 (SD = .29,

Min = .07, *Max* = 1.00), wie oben bereits berichtet waren die Items also tendenziell einfach für die Stichprobe, deckten jedoch einen relativ breiten Schwierigkeitsbereich ab. In der siebten Erhebungswelle lag der Wert für die interne Konsistenz bei α = .87. Die mittlere Trennschärfe betrug r_{it} = .34 (*SD* = .11, *Min* = .10, *Max* = .52) und lag somit auch hier für die meisten Items in einem akzeptablen bis guten Bereich. Die mittlere Itemschwierigkeit lag bei *M* = .78 (*SD* = .29, *Min* = .08, *Max* = 1.00), d.h. der Test beinhaltete nur wenige Items, die nur wenige Teilnehmende richtig beantwortet haben. Eine Übersicht über die einzelnen Trennschärfen und Itemschwierigkeiten sowie die Anzahl gültiger Werte pro Item bietet *Tabelle A 2* in Anhang A. Insgesamt waren die interne Konsistenz und die Itemkennwerte nach Ausschluss des Items Nummer 4 in Block O größtenteils zufriedenstellend, wobei mehr schwierigere Items wünschenswert gewesen wären, die im oberen Kompetenzbereich der Teilnehmenden besser differenzieren, um einen Deckeneffekt zu vermeiden.

4.1.3 DGCF-MAT

Zur Prüfung der Testgüte des DGCF-MAT wurde ebenso Cronbachs α als Maß für die interne Konsistenz bzw. Reliabilität des Tests sowie die Schwierigkeit und Trennschärfe der Items ermittelt. Die interne Konsistenz des DGCF-MAT lag für 1 506 Fälle, für die ein gültiger Summenwert für den gesamten DGCF-MAT berechnet werden konnte, in der ersten Erhebungswelle bei Cronbachs α = .72. Die Trennschärfe lag durchschnittlich bei r_{it} = .36 (*SD* = .08, *Min* = .16, *Max* = .45) und war somit überwiegend akzeptabel bis gut. Die Itemschwierigkeit lag durchschnittlich bei *M* = .55 (*SD* = .18, *Min* = .31, *Max* = .82) und deckte den Kompetenzbereich der Teilnehmenden somit gut ab. Eine Übersicht über die einzelnen Trennschärfen und Itemschwierigkeiten bietet *Tabelle A 3* in Anhang A. Insgesamt sind die Ergebnisse der Skalenanalyse des DGCF-MAT zufriedenstellend. Die Testgüte wurde nur für die erste Erhebungswelle geprüft, da nur die Daten des DGCF-MAT der ersten Erhebungswelle in den folgenden Analysen verwendet werden.

4.1.4 Can-Do-Statements

Wie in Abschnitt 3.3.1.4 beschrieben, sollte aus den Can-Do-Statements zum Verstehen und Sprechen der deutschen Sprache ein Summenwert gebildet werden. Zuerst wurden die interne Konsistenz der Skala und die Itemkennwerte untersucht und die Eindimensionalität als Voraussetzung zur Berechnung des Summenwerts geprüft. Anhand der Ergebnisse und inhaltlicher Überlegungen wurde eine Itemauswahl getroffen, welche erneut geprüft wurde.

Die interne Konsistenz der acht Can-Do-Statements zum Verstehen und Sprechen der deutschen Sprache lag in der ersten Erhebungswelle bei Cronbachs α = .78. Die Trennschärfe lag durchschnittlich bei r_{it} = .50 (*SD* = .09, *Min* = .32, *Max* = .61). Die Itemschwierigkeit lag durchschnittlich

bei $M = .67$ ($SD = .27$, $Min = .29$, $Max = .93$). In der siebten Erhebungswelle lag die interne Konsistenz der acht Can-Do-Statements zum Verstehen und Sprechen der deutschen Sprache bei Cronbachs $\alpha = .82$. Die Trennschärfe lag durchschnittlich bei $r_{it} = .58$ ($SD = .08$, $Min = .48$, $Max = .67$). Die Itemschwierigkeit lag durchschnittlich bei $M = .81$ ($SD = .15$, $Min = .61$, $Max = .94$).

Der Vergleich zwischen einem an die Daten angepassten einfaktoriellen Modell mit einem zweifaktoriellen Modell wies in beiden Erhebungswellen auf eine bessere Passung des zweifaktoriellen Modells hin (s. *Tabelle 6*), weshalb die Annahme der Eindimensionalität der Skala nicht haltbar war und diese Voraussetzung für die Berechnung eines Summenwerts nicht erfüllt war. Deshalb wurde die Skala weiter untersucht. In der ersten Erhebungswelle waren die Ergebnisse des Vergleichs zwischen einem zwei- und einem dreifaktoriellen Modell gemischt, wobei der BIC-Wert für das zweifaktorielle Modell geringer war und dieses sparsamere Modell deshalb bevorzugt wurde. Der Vergleich des zweifaktoriellen mit dem dreifaktoriellen Modell in der siebten Erhebungswelle wies ebenso auf eine bessere Passung des zweifaktoriellen Modells hin. Deshalb wurden die Faktorladungen des zweifaktoriellen Modells in beiden Erhebungswellen genauer betrachtet. Diese sind in *Tabelle B 1* in Anhang B enthalten. In der ersten Erhebungswelle luden die ersten fünf Statements zum Verstehen und Sprechen hoch auf dem zweiten Faktor, das sechste Statement lud auf beiden Faktoren hoch, das siebte Statement lud überwiegend auf dem ersten Faktor, zeigte jedoch zumindest eine geringe Ladung auf dem zweiten Faktor und das letzte Statement lud negativ auf dem zweiten Faktor. In der siebten Erhebungswelle war das Ladungsmuster ähnlich. Hier luden die ersten sechs Statements höher auf dem zweiten Faktor, das siebte Statement lud auf beiden Faktoren und das letzte Statement lud hauptsächlich auf dem ersten Faktor. Da insbesondere das letzte Statement nicht mit den anderen Statements auf demselben Faktor zu laden scheint und sich zudem inhaltlich von den anderen Statements unterscheidet, indem es deutlich länger und komplexer formuliert ist, wurde es ausgeschlossen und die kürzere Skala mit sieben Items erneut evaluiert. Das sechste und siebte Statement zum Verstehen und Sprechen wurden trotz niedriger Ladungen auf dem relevanten Faktor beibehalten, da die anderen Statements insbesondere in der siebten Erhebungswelle hohe Schwierigkeitswerte aufwiesen, also von sehr vielen Teilnehmenden gewählt wurden, und die Skala so zumindest noch zwei etwas schwierigere Items enthielt.

Tabelle 6. Vergleiche ein- und mehrfaktorieller Modelle der explorativen Faktorenanalyse zur Prüfung der Eindimensionalität der Can-Do-Statements in beiden Erhebungswellen

Modell	AIC	BIC	Log-Likelihood	ΔX^2	Δdf	Δp
Erhebungswelle 1						
Can-Do-Statements zum Verstehen und Sprechen (alle Items)						

Modell	AIC	BIC	Log-Like-lihood	ΔX^2	Δdf	Δp
1-Faktor-Modell	11 504.98	11 593.52	–5 736.488			
2-Faktoren-Modell	11 256.28	11 383.56	–5 605.141	262.695	7	< .001
3-Faktoren-Modell	11 255.08	11 415.55	–5 598.538	13.205	6	.04
Can-Do-Statements zum Verstehen und Sprechen (ohne letztes Item)						
1-Faktor-Modell	9 500.67	9 578.15	–4 736.337			
2-Faktoren-Modell	9 416.87	9 527.55	–4 688.437	95.799	6	< .001
Erhebungswelle 7						
Can-Do-Statements zum Verstehen und Sprechen (alle Items)						
1-Faktor-Modell	4 012.23	4 086.72	–1 990.116			
2-Faktoren-Modell	3 979.02	4 086.10	–1 966.511	47.209	7	< .001
3-Faktoren-Modell	3 987.27	4 122.28	–1 964.634	3.753	6	.71
Can-Do-Statements zum Verstehen und Sprechen (ohne letztes Item)						
1-Faktor-Modell	3 179.60	3 244.77	–1 575.798			
2-Faktoren-Modell	3 185.92	3 279.03	–1 572.958	5.678	6	.46

Die interne Konsistenz der sieben ausgewählten Can-Do-Statements zum Verstehen und Sprechen lag in der ersten Erhebungswelle bei Cronbachs $\alpha = .78$. Die Trennschärfe lag durchschnittlich bei $r_{it} = .53$ ($SD = .10$, $Min = .38$, $Max = .65$) und war somit für alle Items gut. Die Itemschwierigkeit lag durchschnittlich bei $M = .73$ ($SD = .24$, $Min = .39$, $Max = .93$), d.h. mit den Statements wurden tendenziell aus Sicht der Teilnehmenden einfache Fähigkeiten beschrieben und insbesondere der obere Kompetenzbereich wurde nicht gut abgedeckt. In der siebten Erhebungswelle lag die interne Konsistenz der sieben ausgewählten Can-Do-Statements zum Verstehen und Sprechen bei Cronbachs $\alpha = .82$. Die Trennschärfe lag durchschnittlich bei $r_{it} = .60$ ($SD = .11$, $Min = .42$, $Max = .70$) und war somit auch in dieser Erhebungswelle für alle Items gut. Die Itemschwierigkeit lag durchschnittlich bei $M = .83$ ($SD = .13$, $Min = .64$, $Max = .94$), d.h. dass es kein Item gab, das gut zwischen Teilnehmenden differenzieren konnte, die ihre Deutschkompetenzen vergleichsweise hoch einschätzten. Eine Übersicht über die einzelnen Trennschärfen und Itemschwierigkeiten für die Skalen mit acht und mit sieben Can-Do-Statements zum Verstehen und Sprechen der deutschen Sprache in der ersten und siebten Erhebungswelle bietet *Tabelle A 4* in Anhang A.

Beim Vergleich eines einfaktoriellen Modells mit einem zweifaktoriellen Modell wurde anhand der Daten der ersten Erhebungswelle das zweifaktorielle und anhand der Daten der siebten Erhebungswelle das einfaktorielle Modell bevorzugt (s. *Tabelle 6*). In der ersten Erhebungswelle scheint also auch für die sieben ausgewählten Can-Do-Statements zum Verstehen und Sprechen keine Eindimensionalität gegeben zu sein. Da die Auswahl jedoch auf theoretischen Überlegungen

basiert und dieselbe Skala in der siebten Erhebungswelle eindimensional ist, wurde dennoch für beide Erhebungswellen ein Summenscore dieser sieben Can-Do-Statements für die weiteren Analysen berechnet.

4.1.5 Nutzung von Möglichkeiten zum Deutschlernen

Um die Items zur Nutzung von Möglichkeiten zum Deutschlernen zu einem Summenwert verrechnen und für die weiteren Analysen nutzen zu können, wurden diese hinsichtlich ihrer internen Konsistenz, Itemkennwerte und Eindimensionalität untersucht und anhand der Ergebnisse und inhaltlicher Überlegungen eine Itemauswahl getroffen, welche erneut geprüft wurde.

Die interne Konsistenz der Items zur Nutzung von Möglichkeiten zum Deutschlernen lag bei Cronbachs $\alpha = .62$. Die Trennschärfe lag durchschnittlich bei $r_{it} = .32$ ($SD = .14$, $Min = .11$, $Max = .48$). Die Itemschwierigkeit lag durchschnittlich bei $M = .47$ ($SD = .19$, $Min = .16$, $Max = .70$).

In *Tabelle 7* ist ersichtlich, dass beim Vergleich der Anpassung der Daten an ein einfaktorielles und an ein zweifaktorielles Modell das zweifaktorielle Modell bevorzugt wurde. Gegenüber dem zweifaktoriellen Modell wurde das dreifaktorielle Modell bevorzugt. Beim Vergleich zwischen der Anpassung der Daten an ein vierfaktorielles Modell und an ein dreifaktorielles Modell waren die Ergebnisse nicht eindeutig, da zumindest das BIC das dreifaktorielle Modell bevorzugte. Deshalb wurden die Faktorladungen des sparsameren dreifaktoriellen Modells untersucht. Es zeigte sich, dass insbesondere die Items 2 bis 6 auf demselben Faktor luden (s. *Tabelle B 2* in Anhang B). Diese Gruppierung scheint auch inhaltlich sinnvoll, da es sich bei diesen fünf Items um Möglichkeiten zum Deutschlernen handelt, die aus eigenem Engagement verfolgt werden können, ohne auf die Hilfe anderer Personen angewiesen zu sein. Die Items 1 und 7 hängen neben dem eigenen Engagement auch von dem sozialen Umfeld der Personen ab. Das achte Item lautet „andere Möglichkeiten“ und ist somit unspezifisch. Da der empirisch gefundene Faktor, der die Items 2 bis 6 beinhaltet, auch aus inhaltlicher Sicht sehr gut zu dem für die weiteren Analysen zu messenden Konstrukt des Engagements beim Deutschlernen passt, wurde die auf die Items 2 bis 6 gekürzte Skala der Nutzung von Möglichkeiten zum Deutschlernen weiter untersucht.

Die interne Konsistenz der fünf ausgewählten Items zur Nutzung von Möglichkeiten zum Deutschlernen lag bei Cronbachs $\alpha = .68$ und somit etwas höher als die interne Konsistenz der gesamten Skala. Die Trennschärfe lag durchschnittlich bei $r_{it} = .44$ ($SD = .05$, $Min = .37$, $Max = .51$) und war für alle Items gut. Die Itemschwierigkeit lag durchschnittlich bei $M = .52$ ($SD = .16$, $Min = .34$, $Max = .70$) und deckte somit insbesondere den mittleren Schwierigkeitsbereich gut ab. Eine Übersicht über die einzelnen Trennschärfen und Itemschwierigkeiten aller acht Items zur

Nutzung von Möglichkeiten zum Deutschlernen und der auf fünf Items gekürzten Skala bietet *Tabelle A 5* in Anhang A.

Die Ergebnisse zum Vergleich eines zweifaktoriellen Modells mit dem einfaktoriellen Modell waren nicht eindeutig (s. *Tabelle 7*). Der BIC-Wert war jedoch für das einfaktorielle Modell geringer, was Grund zur Bevorzugung des sparsameren einfaktoriellen Modells gibt und somit die Eindimensionalität angenommen und ein Summenwert für das Engagement beim Deutschlernen aus den fünf Items berechnet werden kann. Für die weiteren Analysen wurde deshalb der Summenwert aus den Items 2 bis 6 zur Nutzung von Möglichkeiten zum Deutschlernen (s. *Abbildung 9*) als Indikator des Engagements beim Deutschlernen berechnet.

Tabelle 7. Vergleiche ein- und mehrfaktorieller Modelle der explorativen Faktorenanalyse zur Prüfung der Eindimensionalität der Items zur Nutzung von Möglichkeiten zum Deutschlernen in der ersten Erhebungswelle

Modell	AIC	BIC	Log-Likelihood	ΔX^2	Δdf	Δp
8 Items zur Nutzung von Möglichkeiten zum Deutschlernen						
1-Faktor-Modell	17 395.38	17 483.91	–8 681.688			
2-Faktoren-Modell	17 295.43	17 422.71	–8 624.716	113.944	7	< .001
3-Faktoren-Modell	17 253.16	17 413.64	–8 597.580	54.271	6	< .001
4-Faktoren-Modell	17 228.64	17 416.78	–8 580.319	34.522	5	< .001
5 Items zur Nutzung von Möglichkeiten zum Deutschlernen						
1-Faktor-Modell	10 944.26	10 999.59	–5 462.128			
2-Faktoren-Modell	10 933.00	11 010.47	–5 452.501	19.255	4	.001

4.2 Deskriptive Statistiken

Die deskriptiven Statistiken der in den Analysen dieser Arbeit verwendeten Variablen werden in den folgenden Abschnitten 4.2.1 bis 4.2.3 wiedergegeben und für die Selbsteinschätzungen sowie Deutschkompetenztests werden Häufigkeitsverteilungen grafisch dargestellt.

4.2.1 Selbsteinschätzungen

Tabelle 8 enthält die deskriptiven Statistiken der Selbsteinschätzungsitems der Analysestichprobe 1 in der ersten Erhebungswelle. *Tabelle 9* enthält die deskriptiven Statistiken der Selbsteinschätzungsitems der Analysestichprobe 2 in der siebten Erhebungswelle. In *Abbildung 13* ist die

Häufigkeitsverteilung des Standard-Items zum Verstehen der deutschen Sprache in der ersten Erhebungswelle dargestellt und in *Abbildung 14* ist die Häufigkeitsverteilung des Standard-Items zum Verstehen der deutschen Sprache in der siebten Erhebungswelle dargestellt. In *Abbildung 15* und *Abbildung 16* sind die Häufigkeitsverteilungen des Schieberegler-Items und des Vergleich-Items jeweils zum Verstehen der deutschen Sprache dargestellt. Für diese Selbsteinschätzungen sind die Häufigkeitsverteilungen jeweils nur für die Analysestichprobe 1 in der ersten Erhebungswelle dargestellt, da die Items in der siebten Erhebungswelle nicht erhoben wurden. In *Abbildung 17* und in *Abbildung 18* sind die Häufigkeitsverteilungen des Summenwerts der sieben Can-Do-Statements zum Verstehen und Sprechen der deutschen Sprache in der ersten und der siebten Erhebungswelle dargestellt.

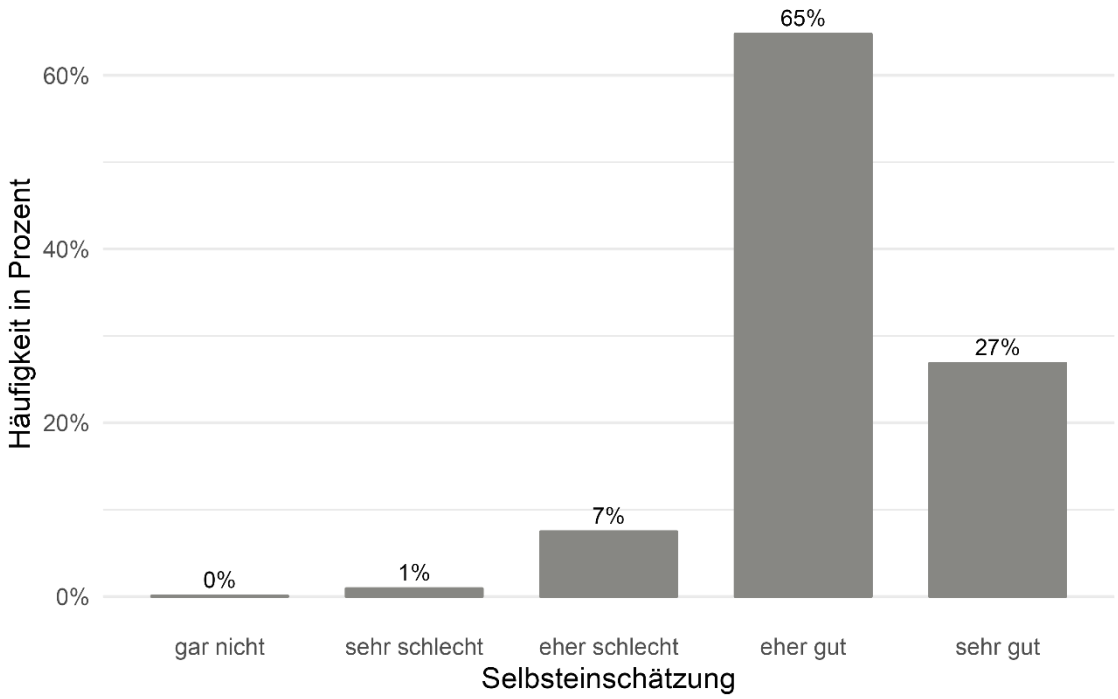
Tabelle 8. Deskriptive Statistiken der Selbsteinschätzungsitems der Analysestichprobe 1 in der ersten Erhebungswelle

	n	*M*	*SD*	Min	Max	Schiefe	Kurtosis
Standard Verstehen	1 877	4.17	0.59	1	5	–0.39	1.18
Standard Sprechen	1 877	4.10	0.62	1	5	–0.52	1.33
Standard Verstehen und Sprechen Mittelwert	1 877	4.14	0.58	1	5	–0.51	1.49
Schieberegler Verstehen	944	7.37	1.59	0	10	–0.85	1.26
Schieberegler Verstehen und Sprechen Mittelwert	944	7.33	1.55	1	10	–0.79	0.86
Vergleich Verstehen	922	3.50	0.76	1	5	–0.45	0.02
Vergleich Verstehen und Sprechen Mittelwert	924	3.49	0.73	1	5	–0.40	–0.20
Can-Do-Statements Verstehen und Sprechen Summenwert (7 Items)	1 870	5.08	1.79	0	7	–1.12	0.77

Tabelle 9. Deskriptive Statistiken der Selbsteinschätzungsitems der Analysestichprobe 2 in der siebten Erhebungswelle

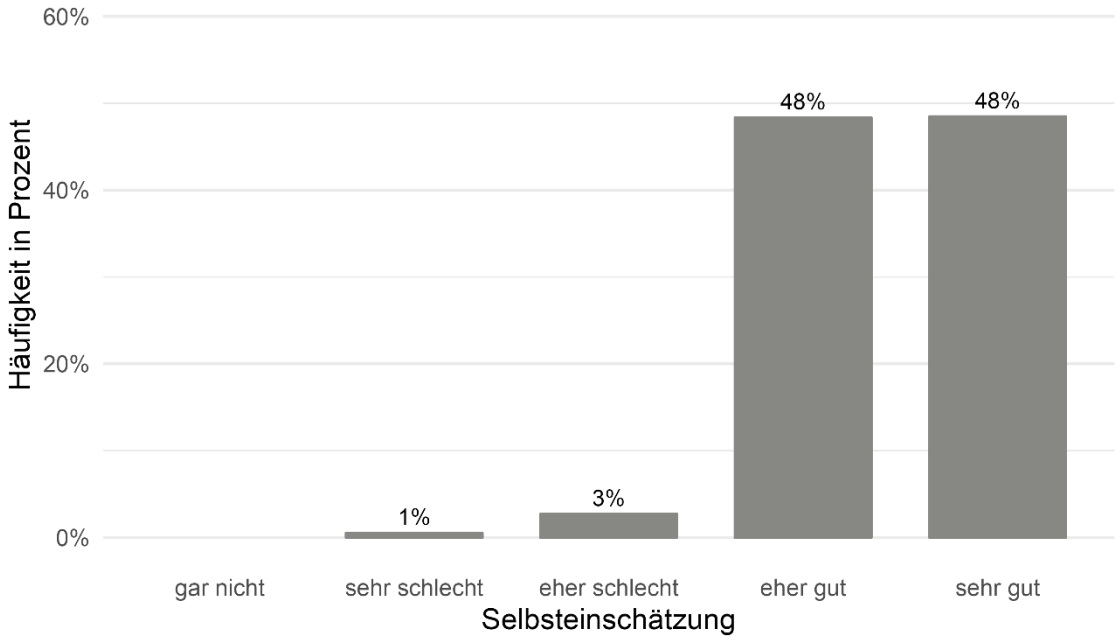
	n	*M*	*SD*	Min	Max	Schiefe	Kurtosis
Standard Verstehen	778	4.45	0.58	2	5	–0.62	0.32
Standard Sprechen	777	4.28	0.59	2	5	–0.48	0.97
Can-Do-Statements Verstehen und Sprechen Summenwert (7 Items)	777	5.85	1.69	0	7	–1.95	3.34

Abbildung 13. Häufigkeitsverteilung der Standard-Selbsteinschätzung zum Verstehen der deutschen Sprache in der ersten Erhebungswelle



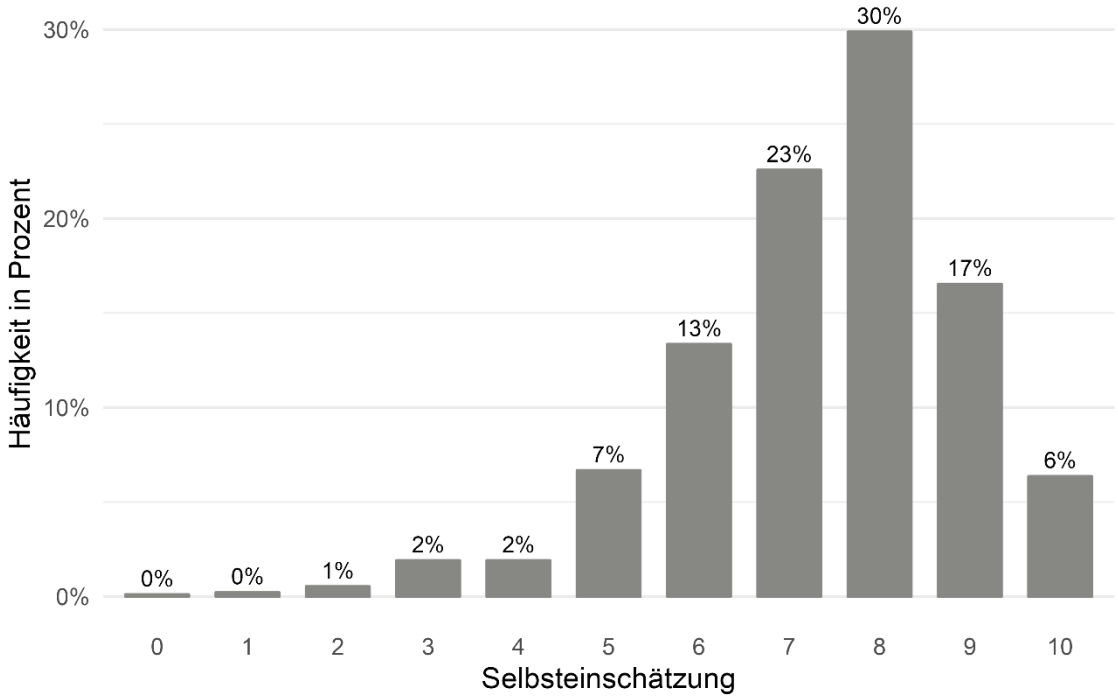
Anmerkungen. $N = 1\,877$.

Abbildung 14. Häufigkeitsverteilung der Standard-Selbsteinschätzung zum Verstehen der deutschen Sprache in der siebten Erhebungswelle



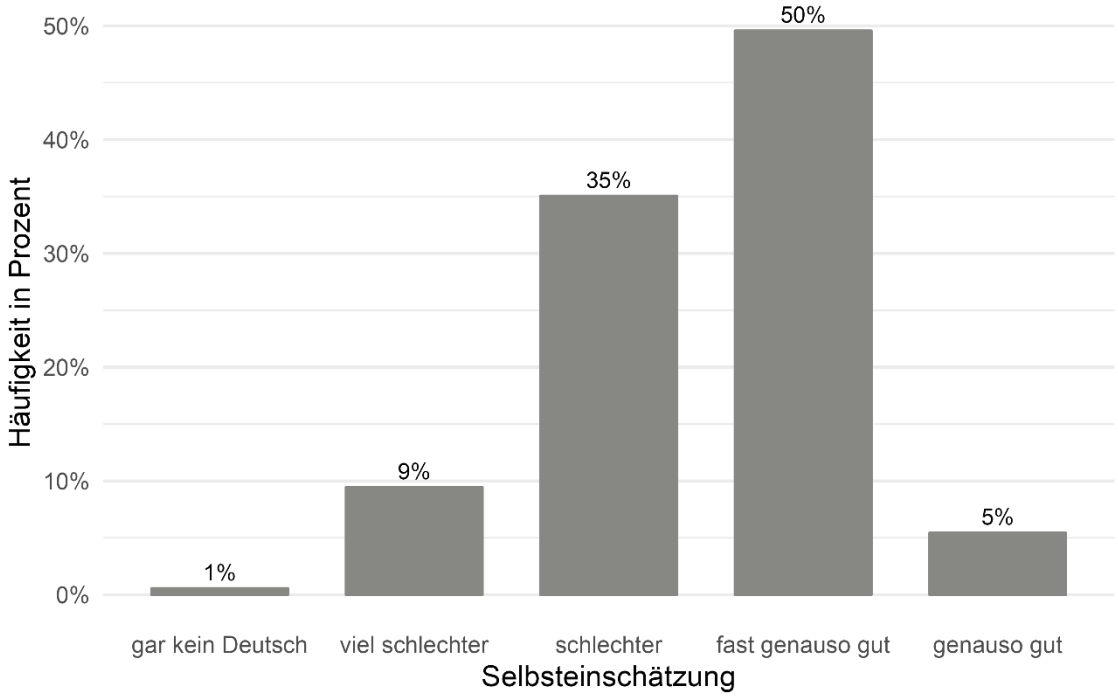
Anmerkungen. $N = 778$.

Abbildung 15. Häufigkeitsverteilung der Schieberegler-Selbsteinschätzung zum Verstehen der deutschen Sprache in der ersten Erhebungswelle



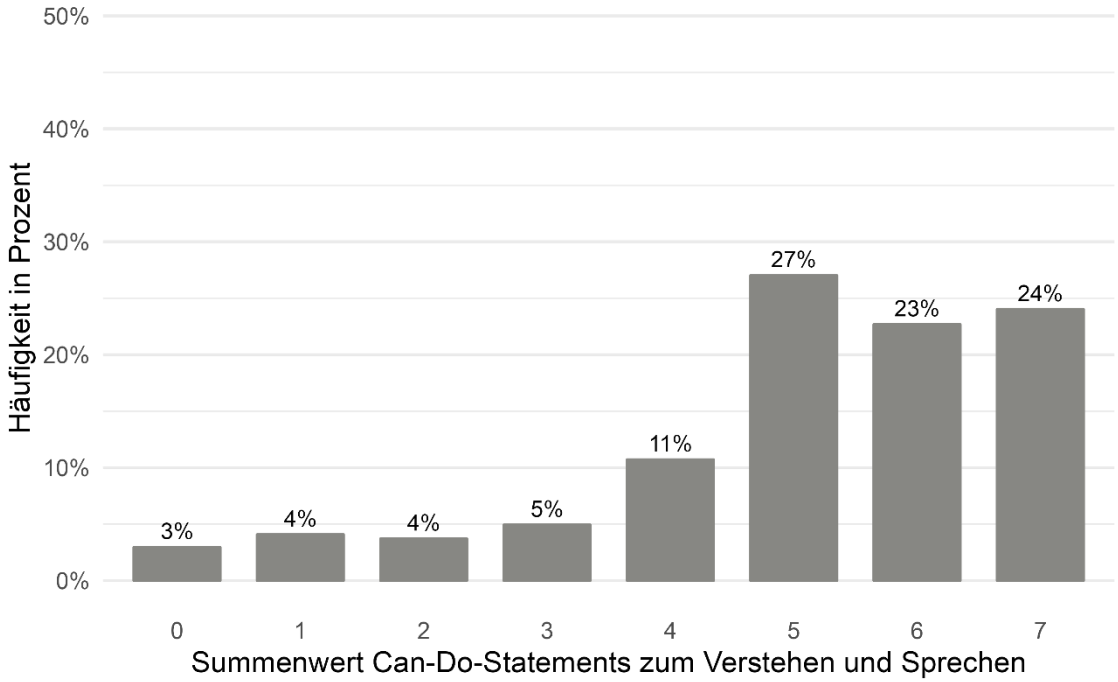
Anmerkungen. $n = 944$.

Abbildung 16. Häufigkeitsverteilung der Vergleich-Selbsteinschätzung zum Verstehen der deutschen Sprache in der ersten Erhebungswelle



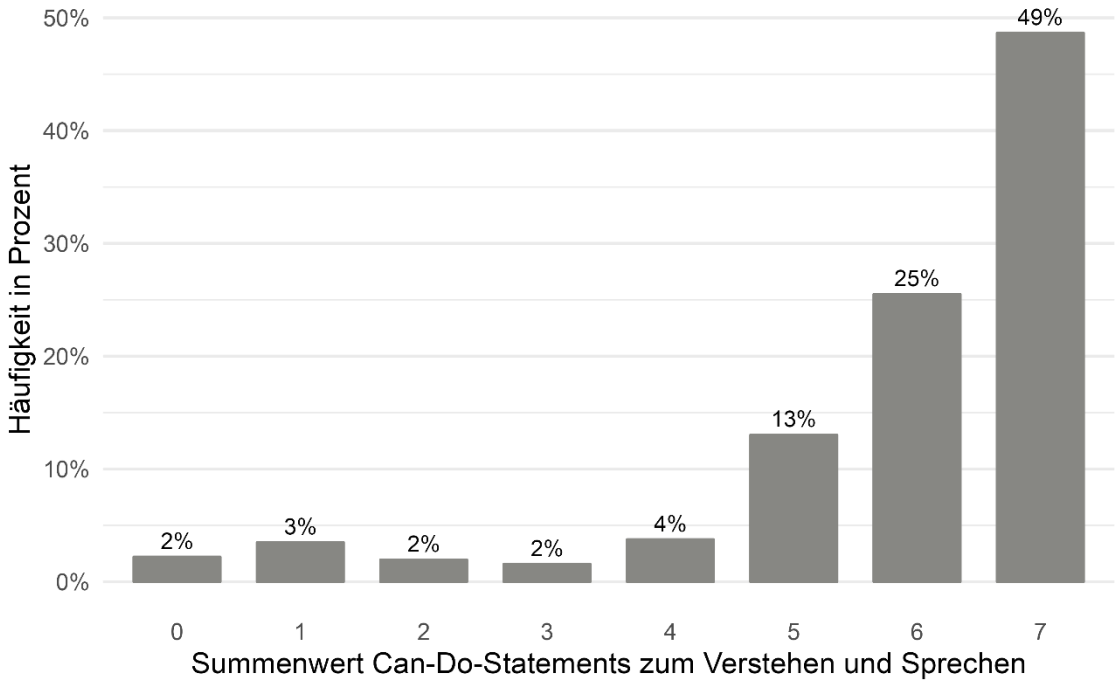
Anmerkungen. $n = 922$.

Abbildung 17. Häufigkeitsverteilung des Summenwerts der sieben Can-Do-Statements zum Verstehen und Sprechen der deutschen Sprache in der ersten Erhebungswelle



Anmerkungen. $n = 1\,870$.

Abbildung 18. Häufigkeitsverteilung des Summenwerts der sieben Can-Do-Statements zum Verstehen und Sprechen der deutschen Sprache in der siebten Erhebungswelle



Anmerkungen. $n = 777$.

4.2.2 Deutschkompetenztests

Tabelle 10 enthält deskriptive Statistiken zu den Deutschkompetenztests der Analysestichprobe 1 in der ersten Erhebungswelle. *Tabelle 11* enthält die entsprechenden Statistiken für die Analysestichprobe 2 in der siebten Erhebungswelle. In *Abbildung 19* und in *Abbildung 20* ist die Verteilung des Summenwerts des PPVT-4 in der ersten und der siebten Erhebungswelle grafisch dargestellt. In *Abbildung 21* und in *Abbildung 22* ist die Verteilung des Summenwerts des TROG-D in der ersten und der siebten Erhebungswelle dargestellt.

Tabelle 10. Deskriptive Statistiken der Deutschkompetenztests der Analysestichprobe 1 in der ersten Erhebungswelle

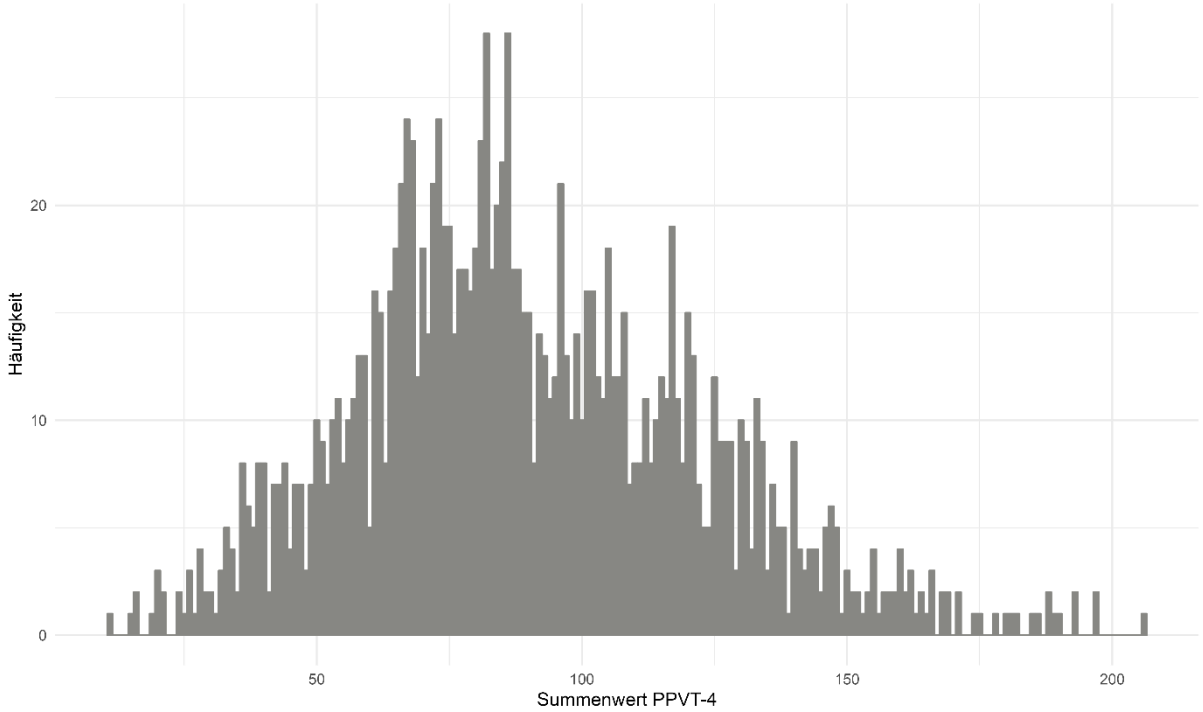
	n	*M*	*SD*	Min	Max	Schiefe	Kurtosis
PPVT-4 Summenwert	1 391	89.73	32.35	11	206	0.46	0.16
PPVT-4 WLE	1 449	0.11	1.78	−6.10	6.14	−0.04	0.19
TROG-D Summenwert (47 Items)	1 801	33.48	5.87	9	47	−0.73	0.57
Kombinierter Deutschkompetenzscore	1 877	3.50	1.00	−0.54	6.33	−0.45	0.43

Anmerkungen. WLE = Weighted Likelihood Estimate.

Tabelle 11. Deskriptive Statistiken der Deutschkompetenztests der Analysestichprobe 2 in der siebten Erhebungswelle

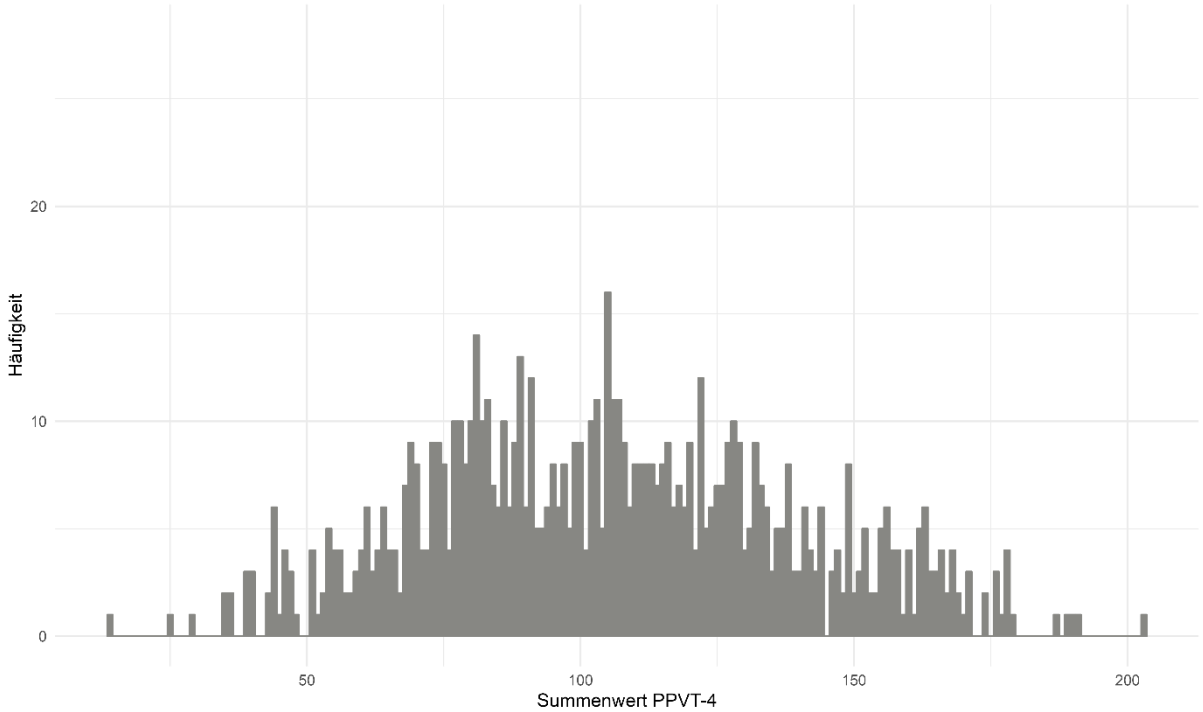
	n	*M*	*SD*	Min	Max	Schiefe	Kurtosis
PPVT-4 Summenwert	773	105.45	33.27	14	203	0.18	–0.47
TROG-D Summenwert (47 Items)	744	36.29	5.55	1	47	–1.43	4.23
Kombinierter Deutschkompetenzscore	778	3.50	1.00	–3.13	6.06	–0.76	2.74

Abbildung 19. Häufigkeitsverteilung des Summenwerts des PPVT-4 in der ersten Erhebungswelle



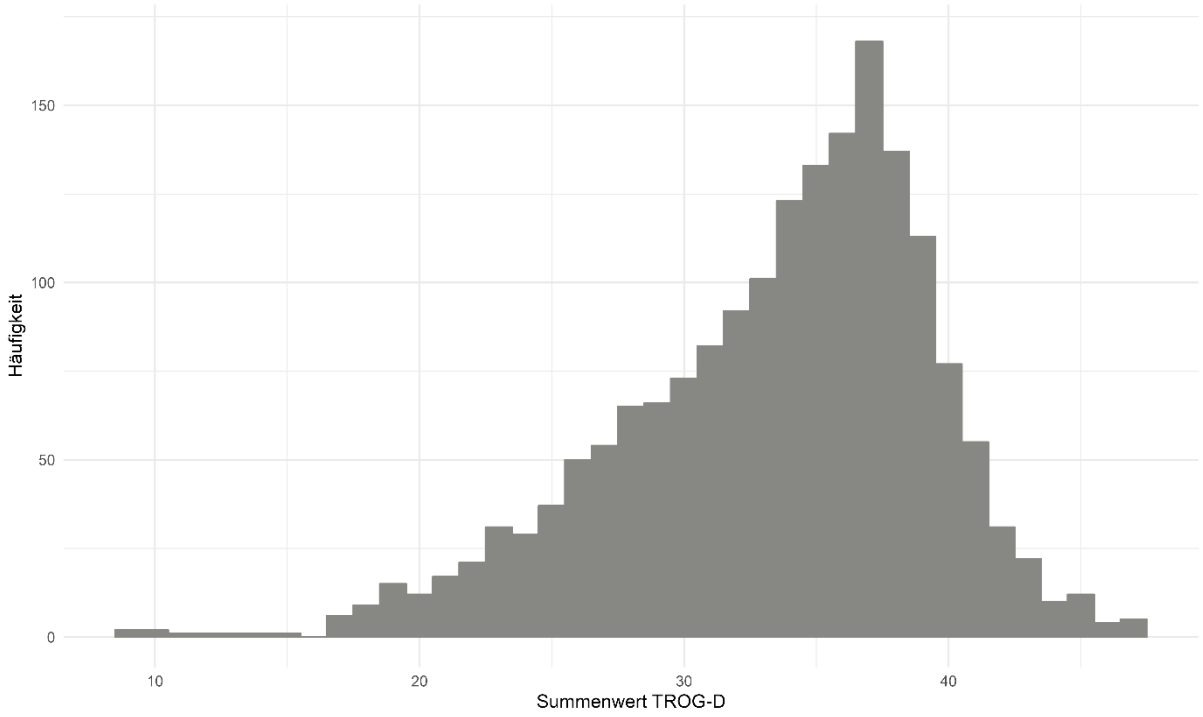
Anmerkungen. $n = 1\,391$.

Abbildung 20. Häufigkeitsverteilung des Summenwerts des PPVT-4 in der siebten Erhebungswelle



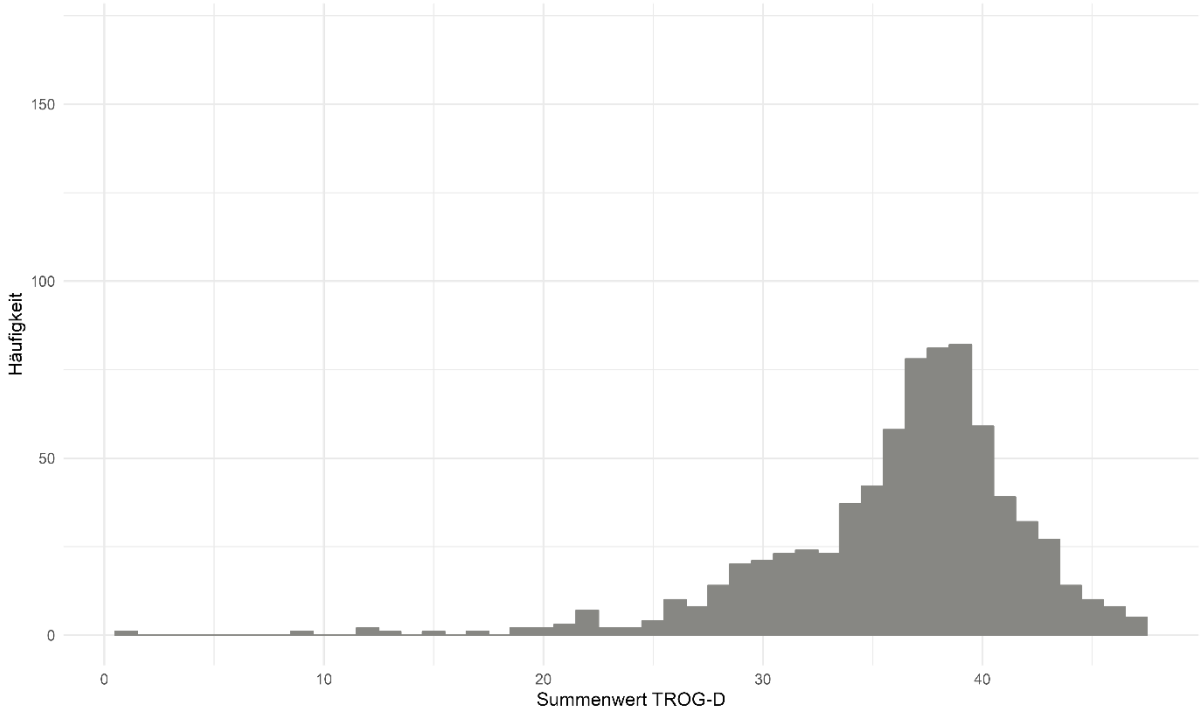
Anmerkungen. $n = 773$.

Abbildung 21. Häufigkeitsverteilung des Summenwerts der 47 verwendeten Items des TROG-D in der ersten Erhebungswelle



Anmerkungen. $n = 1\,801$.

Abbildung 22. Häufigkeitsverteilung des Summenwerts der 47 verwendeten Items des TROG-D in der siebten Erhebungswelle



Anmerkungen. $n = 744$.

4.2.3 Weitere Variablen

In *Tabelle 12* sind die deskriptiven Statistiken zu den Variablen enthalten, die für die Prüfung der Hypothesen 2a bis 2e zu den Einflussfaktoren von Selbsteinschätzungen verwendet werden.

Tabelle 12. Deskriptive Statistiken weiterer Variablen der Analysestichprobe 1 in der ersten Erhebungswelle

	n	*M*	*SD*	Min	Max	Schiefe	Kurtosis
Leistung in Mathematik (Selbstauskunft)	1 693	–2.63	1.19	–6	–1	–0.64	0.02
Leistung in Mathematik (Auskunft Lehrkraft)	233	–3.28	1.33	–6	–1	–0.29	–0.66
DGCF-MAT Set 1 Summenwert	1 596	2.22	1.02	0	4	–0.25	–0.36
DGCF-MAT Set 2 Summenwert	1 544	1.84	1.26	0	4	0.17	–1.00
DGCF-MAT Set 3 Summenwert	1 520	2.34	1.20	0	4	–0.29	–0.84
Nutzung Möglichkeiten Deutschlernen Summenwert (5 Items)	1 870	2.59	1.59	0	5	–0.06	–1.07
Teilnahme Deutschkurs	1 868	0.62	0.49	0	1	–0.48	–1.77
Teilnahme Deutschtest	1 865	0.26	0.44	0	1	1.09	–0.82

4.3 Bivariate Zusammenhänge

Tabellen mit Korrelationen zwischen allen Variablen jeder Erhebungswelle sind in Anhang C zu finden. Für die erste Erhebungswelle sind diese in *Tabelle C 1* enthalten und für die siebte Erhebungswelle in *Tabelle C 2*.

4.4 Genauigkeit der Selbsteinschätzungen

In diesem Kapitel werden die Analysen und Ergebnisse zu den Hypothesen zur Genauigkeit der Selbsteinschätzungen beschrieben. Zuerst wird in Abschnitt 4.4.1 auf die Prüfung der Hypothese 1a zur Diskrimination der Selbsteinschätzungen eingegangen, anschließend wird in Abschnitt 4.4.2 auf die Prüfung der Hypothese 1b zur allgemeinen Verzerrung der Selbsteinschätzungen eingegangen und abschließend wird in Abschnitt 4.4.3 auf die Prüfung der Hypothese 1c zur Variation der Selbsteinschätzungen eingegangen. Dabei werden jeweils alle Arten von Selbsteinschätzungsitems betrachtet.

4.4.1 Diskrimination

Die Korrelationen zwischen Selbsteinschätzungen und Kompetenzwerten lagen in der ersten Erhebungswelle für die verschiedenen verwendeten Maße zwischen $r = .24$ und $r = .35$ (s. *Tabelle 13*). In der siebten Erhebungswelle lagen die Korrelationen zwischen den verwendeten Selbsteinschätzungsitems und Kompetenzwerten zwischen $r = .33$ und $r = .39$ (s. *Tabelle 14*). Somit sind die empirisch gefundenen Korrelationen keine wie in Hypothese 1a vorhergesagten Korrelationen geringer Effektstärke sondern sie liegen nach Cohen (1992) in einem Bereich mittlerer Effektstärke.

Tabelle 13. Korrelationen zwischen Selbsteinschätzungen und Ergebnissen der Deutschkompetenztests der Analysestichprobe 1 in der ersten Erhebungswelle

	1	2	3	4	5	6	7
1 Standard Verstehen	1 877	944	922	1 870	1 449	1 801	1 877
2 Schieberegler Verstehen	.59	944	0	941	749	907	944
3 Vergleich Verstehen	.48	-	922	921	689	883	922
4 Can-Do-Statements Verstehen und Sprechen Summenwert (7 Items)	.21	.26	.25	1 870	1 442	1 794	1 870
5 PPVT-4 WLEs	.32	.33	.35	.24	1 449	1 373	1 449
6 TROG-D Summenwert (47 Items)	.28	.28	.29	.25	.72	1 801	1 801
7 Kombinierter Deutschkompetenzscore	.30	.31	.30	.25	.93	.95	1 877

Anmerkungen. Pearson-Korrelationen unterhalb der Diagonalen, Stichprobengrößen der paarweisen Korrelationen oberhalb der Diagonalen. WLE = Weighted Likelihood Estimate.
Alle $p < .001$.

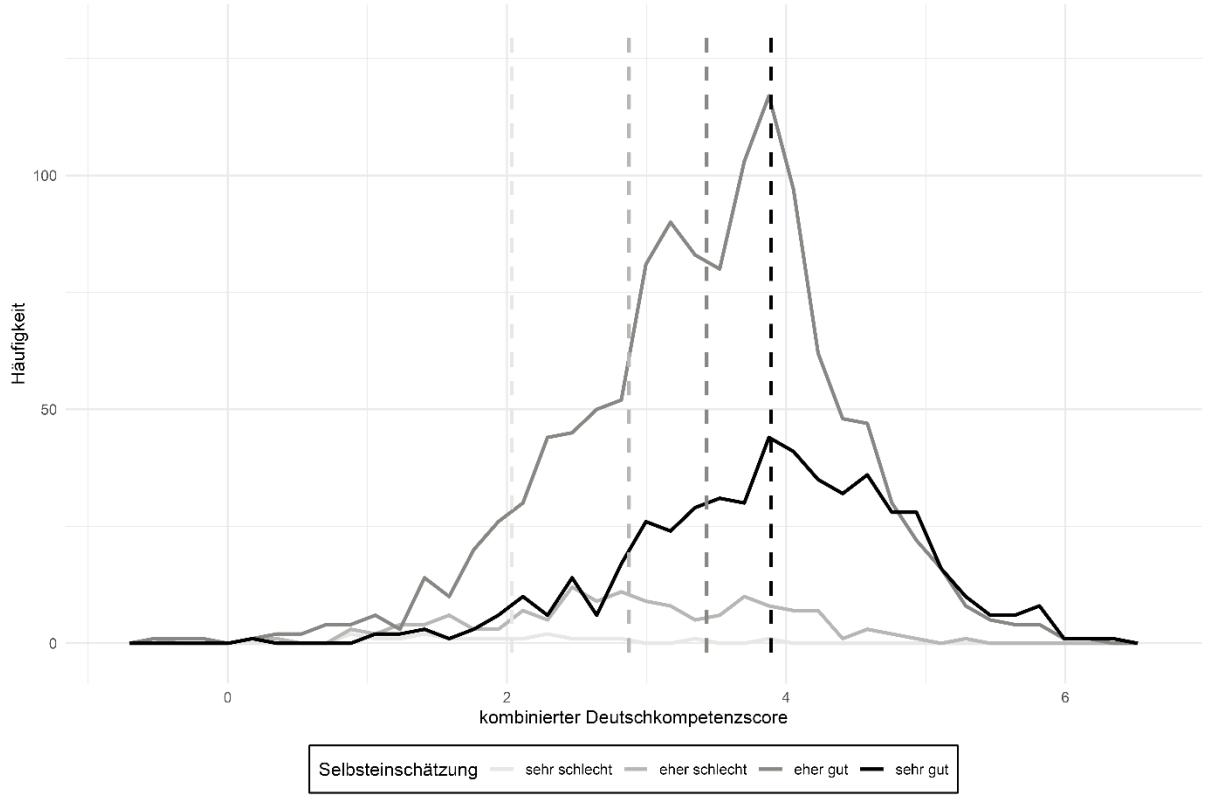
Tabelle 14. Korrelationen zwischen Selbsteinschätzungen und Ergebnissen der Deutschkompetenztests der Analysestichprobe 2 in der siebten Erhebungswelle

	1	2	3	4	5
1 Standard Verstehen	778	777	773	744	778
2 Can-Do-Statements Verstehen und Sprechen Summenwert (7 Items)	.32	777	772	743	777
3 PPVT-4 Summenwert	.33	.36	773	739	773
4 TROG-D Summenwert (47 Items)	.35	.37	.71	744	744
5 Kombinierter Deutschkompetenzscore	.37	.39	.93	.93	778

Anmerkungen. Pearson-Korrelationen unterhalb der Diagonalen, Stichprobengrößen der paarweisen Korrelationen oberhalb der Diagonalen.
Alle $p < .001$.

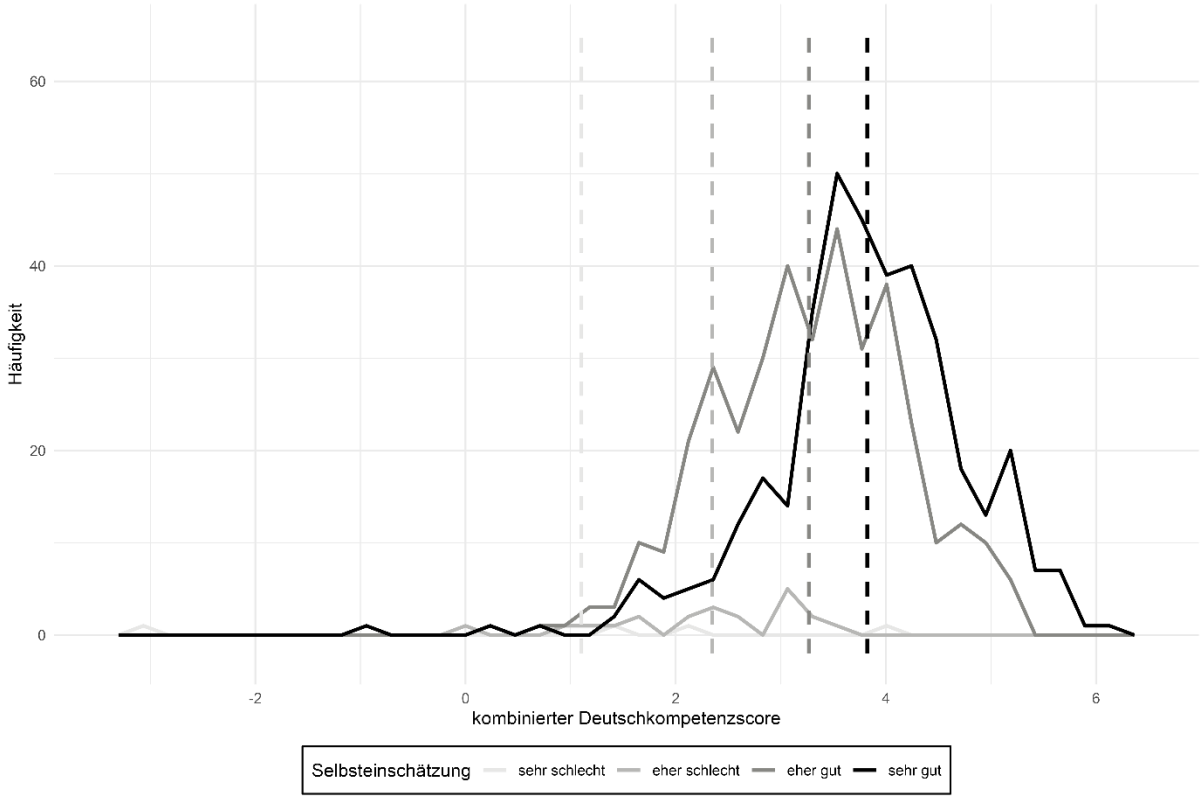
Um die Zusammenhänge zwischen Selbsteinschätzungen und Kompetenzwerten zu veranschaulichen, habe ich für jede Erhebungswelle eine Grafik (s. *Abbildung 23* und *Abbildung 24*) erstellt, die die Verteilung des kombinierten Deutschkompetenzscores nach gewählter Kategorie beim Standard-Item zum Verstehen der deutschen Sprache wiedergibt, sodass innerhalb der Grafik die Häufigkeitsverteilung und der Mittelwert für jede Gruppe von Personen, die dieselbe Antwortkategorie des Selbsteinschätzungsitems gewählt hat, separat in einer unterschiedlichen Graustufe dargestellt wird. Es ist zu erkennen, dass die Mittelwerte des kombinierten Deutschkompetenzscores entsprechend den Selbsteinschätzungskategorien aufsteigend geordnet sind. Die Häufigkeitsverteilungen der verschiedenen Gruppen überlappen sich stark.

Abbildung 23. Häufigkeitsverteilung des kombinierten Deutschkompetenzscores nach Standard-Selbsteinschätzung in der ersten Erhebungswelle



Anmerkungen. Die Antwortkategorie *gar nicht* wurde nicht berücksichtigt. Die gestrichelten Linien kennzeichnen den Mittelwert der Verteilung des Deutschkompetenzscores jeder Gruppe von Personen, die die entsprechende Selbsteinschätzungskategorie gewählt haben. $n_{\text{sehr schlecht}} = 17$; $n_{\text{eher schlecht}} = 140$; $n_{\text{eher gut}} = 1\,215$; $n_{\text{sehr gut}} = 504$; $n_{\text{gesamt}} = 1\,876$.

Abbildung 24. Häufigkeitsverteilung des kombinierten Deutschkompetenzscores nach Standard-Selbsteinschätzung in der siebten Erhebungswelle



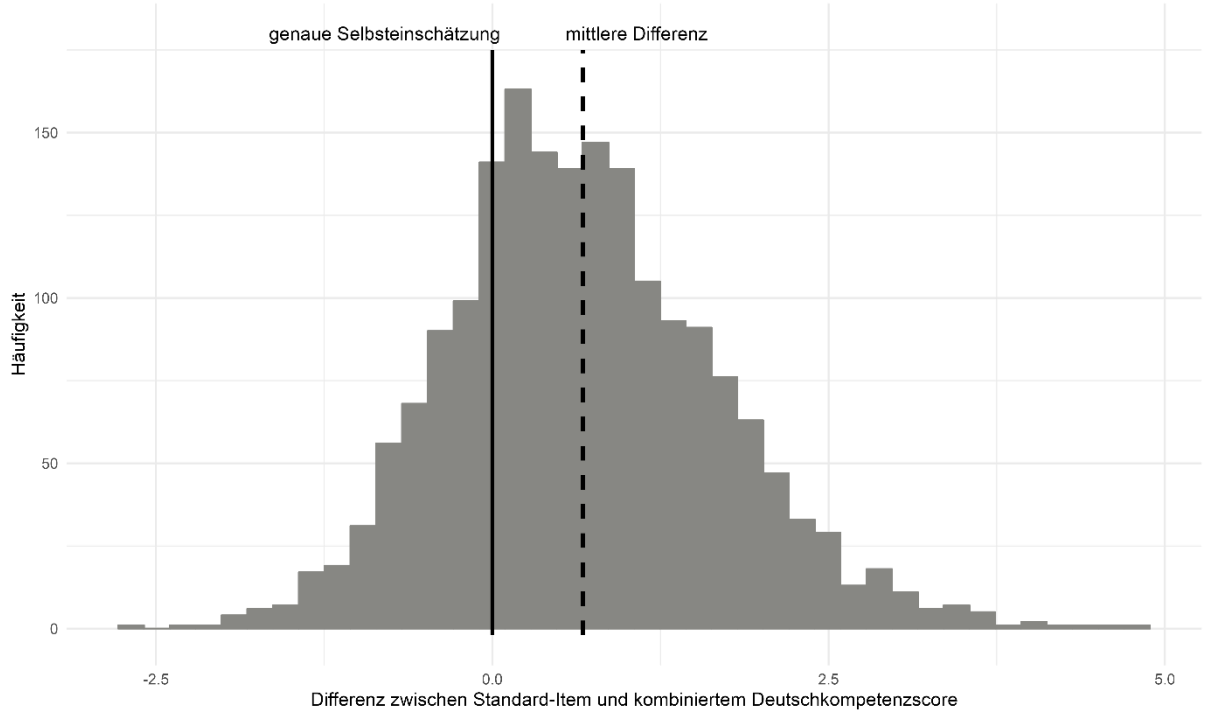
Anmerkungen. Die Antwortkategorie *gar nicht* wurde nicht berücksichtigt. Die gestrichelten Linien kennzeichnen den Mittelwert der Verteilung des Deutschkompetenzscores jeder Gruppe von Personen, die die entsprechende Selbsteinschätzungskategorie gewählt haben. $n_{\text{sehr schlecht}} = 4$; $n_{\text{eher schlecht}} = 21$; $n_{\text{eher gut}} = 376$; $n_{\text{sehr gut}} = 377$; $n_{\text{gesamt}} = 778$.

4.4.2 Allgemeine Verzerrung

Zur Überprüfung der Hypothese 1b, dass die jugendlichen Flüchtlinge ihre Deutschkompetenz im Durchschnitt überschätzen, wurde die Differenz zwischen dem kombinierten Deutschkompetenzscore und dem Standard-Item zur Selbsteinschätzung der Kompetenz im Verstehen der deutschen Sprache untersucht. In Abschnitt 3.3.2.3 ist beschrieben, wie der kombinierte Deutschkompetenzscore so skaliert wurde, dass er der erwarteten Zuordnung der getesteten Kompetenzen zu den Selbsteinschätzungen entspricht, welche auf der fünfstufigen Skala von 1 = *gar nicht* bis 5 = *sehr gut* gemessen wurde. Im Durchschnitt überschätzten die jugendlichen Flüchtlinge ihre Deutschkompetenz. Die Differenz betrug in der Analysestichprobe 1 in der ersten Erhebungswelle im Durchschnitt $M = 0.67$ ($SD = 1.00$, $Min = -2.73$, $Max = 4.75$) und war signifikant größer als 0 ($t(1876) = 29.19$, $p < .001$, $d = 0.67$, 95% KI für d [0.63, ∞]). In der Analysestichprobe 2 in der siebten Erhebungswelle wurde der kombinierte Deutschkompetenzscore auf Basis der Testergebnisse dieser Stichprobe in der siebten Erhebungswelle skaliert. Die Differenz zwischen Standard-

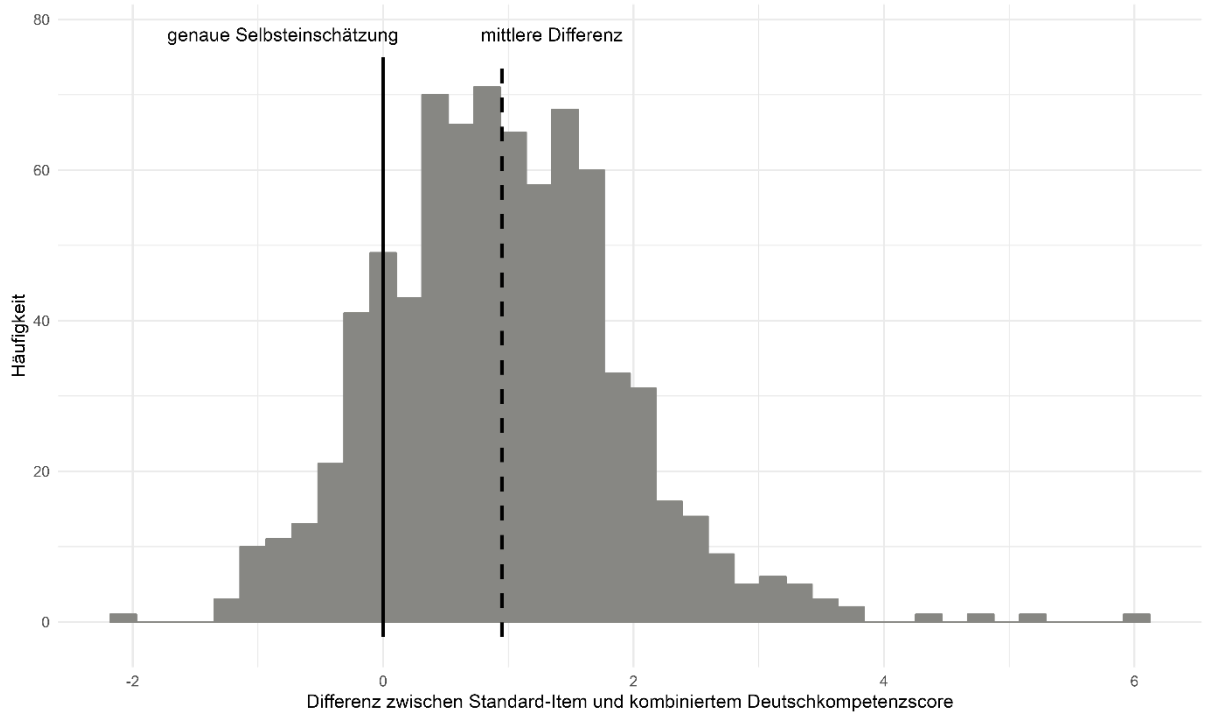
Selbsteinschätzungsitem und kombiniertem Deutschkompetenzscore betrug im Durchschnitt $M = 0.95$ ($SD = 0.95$, $Min = -2.04$, $Max = 6.06$) und war signifikant größer als 0 ($t(777) = 27.74$, $p < .001$, $d = 0.99$, 95% KI für d [0.92, ∞]). Die Verteilung der Differenz zwischen Selbsteinschätzung und kombiniertem Deutschkompetenzscore ist für die Analysestichprobe 1 in der ersten Erhebungswelle in *Abbildung 25* und für die Analysestichprobe 2 in der siebten Erhebungswelle in *Abbildung 26* dargestellt.

Abbildung 25. Histogramm zur Darstellung der Verteilung der Differenz zwischen Selbsteinschätzung und kombiniertem Deutschkompetenzscore in der ersten Erhebungswelle



Anmerkungen. $N = 1\,877$.

Abbildung 26. Histogramm zur Darstellung der Verteilung der Differenz zwischen Selbsteinschätzung und kombiniertem Deutschkompetenzscore in der siebten Erhebungswelle



Anmerkungen. $N = 778$.

Geht man auch im Fall des Schieberegler-Items davon aus, dass entsprechend der erläuterten Definition einer genauen Einschätzung der Mittelwert der Selbsteinschätzungen auf der Mitte der Skala und somit bei 5.5 liegen müsste, falls keine allgemeine Verzerrung vorlag, so überschätzten die $n = 944$ Jugendlichen Flüchtlinge der Analysestichprobe 1, die die Schieberegler-Items beantworteten, ihre Deutschkompetenz auch auf dieser Skala im Durchschnitt ($M = 7.37$, $SD = 1.59$) und der Mittelwert war signifikant größer als 5.5 ($t(943) = 36.15$, $p < .001$, $d = 1.18$, 95% KI für d [1.11, ∞]).

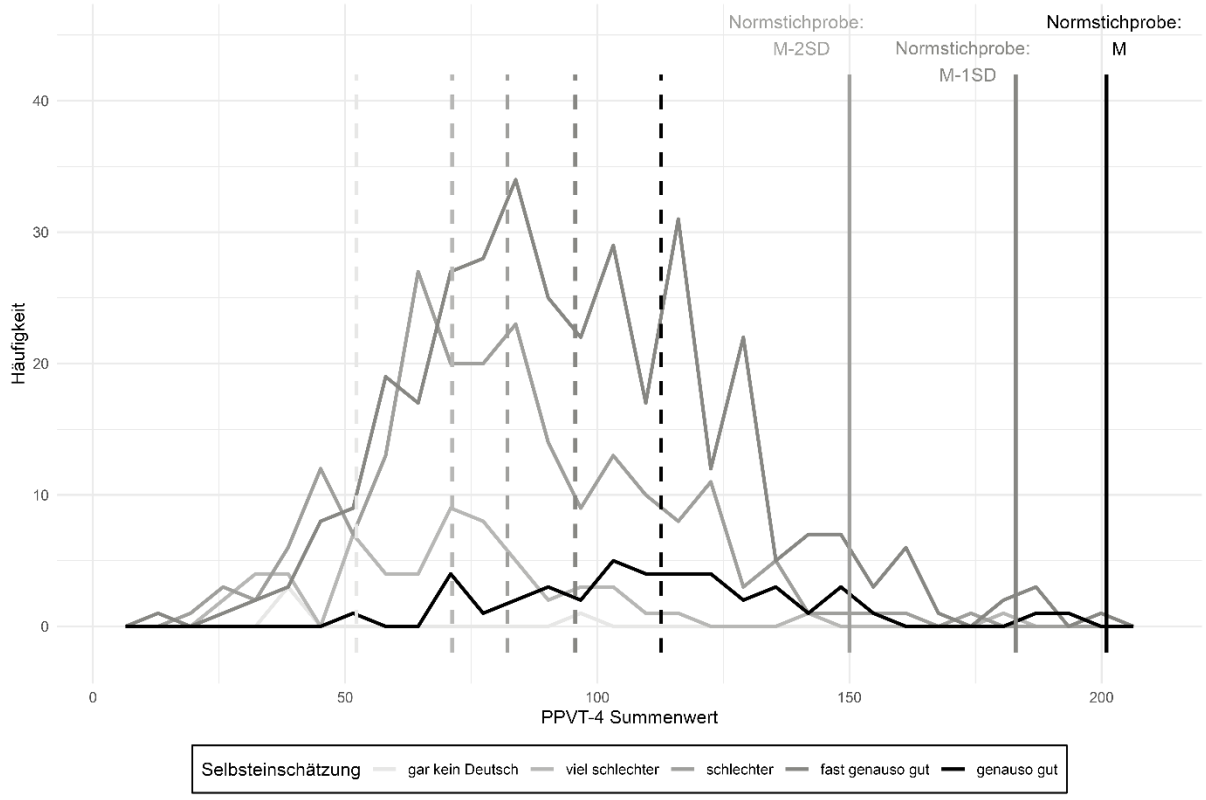
Für die Vergleich-Items können zur Überprüfung der allgemeinen Verzerrung die Normwerte des PPVT-4 herangezogen werden. In der Normstichprobe des PPVT-4, welche überwiegend aus Personen bestand, deren Muttersprache Deutsch war, erreichten Jugendliche im Alter von 16;0 bis 16;5 Jahren durchschnittlich 201 Punkte. Jugendliche, deren Leistung eine Standardabweichung unter dem Mittelwert lag, erreichten ca. 183 Punkte. Jugendliche, deren Leistung zwei Standardabweichungen unter dem Mittelwert lag, erreichten ca. 150 Punkte. Auch hier ist zunächst zu definieren, welche Selbsteinschätzung bei welchem Kompetenzwert einer genauen Selbsteinschätzung entspricht. Der Bereich einer Normalverteilung, der zwischen dem Mittelwert und einer Standardabweichung unter dem Mittelwert liegt, wird üblicherweise als *unterer Normalbereich* bezeichnet. Der Bereich zwischen ein und zwei Standardabweichungen unter dem Mittelwert wird als *unterdurchschnittlich* bezeichnet und der Bereich, der mehr als zwei Standardabweichungen unter

dem Mittelwert liegt wird als *stark unterdurchschnittlich* bezeichnet (z.B. Lenhard et al., 2015, S. 54). Als genaue Selbsteinschätzung wird hier definiert, dass der Kompetenzwert von Personen, die ihre sprachliche Kompetenz als *genauso gut wie ein Deutscher* einstufen, mindestens im unteren Normalbereich liegen muss, d.h. dass deren Summenwert bei min. 183 Punkten liegen muss, damit die Selbsteinschätzung als genau gewertet wird. Für den unterdurchschnittlichen Kompetenzbereich, also 150 bis 183 Summenwertpunkte, wird die Selbsteinschätzung *fast genauso gut wie ein Deutscher* als angemessen gewertet und für den stark unterdurchschnittlichen Kompetenzbereich werden die Selbsteinschätzungen *schlechter als ein Deutscher*, *viel schlechter als ein Deutscher* und *gar nicht* als angemessen gewertet, wobei die Kompetenzen von Personen, die die Kategorie *schlechter als ein Deutscher* gewählt haben, nicht mehr als drei Standardabweichungen vom Mittelwert abweichen sollten, da ansonsten die Kategorien *viel schlechter als ein Deutscher* oder *gar nicht* angemessener sind. Da in der Normwerttabelle des PPVT-4 für Abweichungen von drei oder mehr Standardabweichungen vom Mittelwert keine Summenwerte mehr angegeben werden, kann der für diese Kategorien als angemessen erachtete Wertebereich des Summenwerts nicht spezifiziert werden.

In *Abbildung 27* sind die Verteilungen der Summenwerte nach gewählter Antwortkategorie des Vergleich-Items zum Verstehen der deutschen Sprache dargestellt und deren Mittelwerte gekennzeichnet. Darüber hinaus sind der Mittelwert der ungefähr gleichaltrigen Gruppe der Normstichprobe und die Werte, die ein und zwei Standardabweichungen unter dem Mittelwert dieser Gruppe liegen, eingezeichnet. Die $n = 42$ jugendlichen Flüchtlinge der Analysestichprobe 1, die angaben, Deutsch genauso gut zu sprechen, wie ein Deutscher, erreichten durchschnittlich $M = 112.67$ Punkte auf dem Summenwert des PPVT-4 ($SD = 30.44$) und wichen somit im Durchschnitt deutlich mehr als eine Standardabweichung vom Mittelwert der Normstichprobe ab. Entsprechend der Definition, dass eine genauso gute Leistung maximal eine Standardabweichung vom Mittelwert der Normstichprobe hätte nach unten abweichen dürfen, liegt für diese Gruppe eine deutliche Überschätzung vor. Die $n = 342$ Jugendlichen, die angaben, fast genauso gut Deutsch zu sprechen, wie gleichaltrige Personen mit deutscher Muttersprache, erreichten durchschnittlich $M = 95.62$ ($SD = 30.80$) Punkte auf dem Summenwert des PPVT-4 und wichen ebenfalls mehrere Standardabweichungen vom Mittelwert der Normstichprobe ab, sodass deren Einschätzung ebenfalls als Überschätzung zu werten ist, sofern man davon ausgeht, dass eine fast genauso gute Leistung maximal zwei Standardabweichungen vom Mittelwert der Normstichprobe hätte abweichen dürfen. Die $n = 212$ Jugendlichen, die angaben, schlechter Deutsch zu sprechen als gleichaltrige Personen mit deutscher Muttersprache, erreichten durchschnittlich $M = 82.17$ ($SD = 27.71$) Punkte auf dem Summenwert des PPVT-4. Auch hier ist die Abweichung vom Mittelwert der Normstichprobe so groß, dass diese drei Standardabweichungen wahrscheinlich übersteigt und die Wahl einer niedrigeren Kategorie als angemessener erscheint, weshalb auch diese Angaben im Durchschnitt als

Überschätzung gewertet werden können. Die $n = 59$ Jugendlichen, die angaben, viel schlechter Deutsch zu sprechen als gleichaltrige Personen mit deutscher Muttersprache, erreichten im Durchschnitt $M = 71.22$ ($SD = 27.87$) Punkte beim PPVT-4. Da sie nach dem Ergebnis des PPVT-4 zumindest über geringe Deutschkenntnisse zu verfügen scheinen, ist die Wahl dieser Kategorie angemessen. Gar kein Deutsch zu sprechen gaben nur $n = 4$ Teilnehmende an und erreichten durchschnittlich $M = 52.25$ ($SD = 28.51$) Punkte beim PPVT-4, was tendenziell als Unterschätzung der Deutschkompetenz zu werten ist, da sie zumindest über ein geringes Vokabular im Deutschen zu verfügen scheinen. Diese wenigen Personen fallen jedoch kaum ins Gewicht. Insgesamt haben die jugendlichen Flüchtlinge ihre Deutschkompetenz auf dem Vergleich-Item deutlich überschätzt.

Abbildung 27. Häufigkeitsverteilung des PPVT-4 Summenwerts nach Selbsteinschätzung auf dem Vergleich-Item in der ersten Erhebungswelle



Anmerkungen. Die gestrichelten Linien kennzeichnen den Mittelwert der Verteilung des PPVT-4 Summenwerts jeder Gruppe von Personen, die die entsprechende Selbsteinschätzungskategorie gewählt haben. Die durchgezogenen senkrechten Linien kennzeichnen den Mittelwert der Normstichprobe des PPVT-4 sowie die Werte, die eine bzw. zwei Standardabweichungen vom Mittelwert der Normstichprobe nach unten abweichen (vgl. Lenhard et al., 2015). $n_{\text{gar kein Deutsch}} = 4$; $n_{\text{viel schlechter}} = 59$; $n_{\text{schlechter}} = 212$; $n_{\text{fast genauso gut}} = 342$; $n_{\text{genauso gut}} = 42$; $n_{\text{gesamt}} = 659$.

Für die Can-Do-Statements ist eine Prüfung der allgemeinen Verzerrung der Selbsteinschätzungen schwer möglich, da die Angaben kaum überprüft werden können, weshalb an dieser Stelle darauf verzichtet wird.

Insgesamt deuten die Ergebnisse einheitlich darauf hin, dass die jugendlichen Flüchtlinge ihre Kompetenz im Verstehen der deutschen Sprache durchschnittlich überschätzten, womit Hypothese 1b bestätigt wurde.

4.4.3 Variation

In Hypothese 1c wurde vorhergesagt, dass die Variation der Selbsteinschätzungen geringer sei als die der objektiv gemessenen Kompetenzen. Bei Betrachtung der Standardabweichungen als Maß für die Variation der Werte zeigt sich, dass die Summenwerte der Deutschkompetenztests deutlich stärker variieren als die Werte auf den Selbsteinschätzungsskalen (s. *Tabelle 8*, *Tabelle 9*, *Tabelle 10* und *Tabelle 11*). Jedoch sind die Standardabweichungen abhängig von der Skalierung der Skalen und somit beliebig transformierbar, weshalb im Folgenden weitere Betrachtungen zur Hypothese 1c herangezogen werden.

Ca. 99% der Antworten auf das Standard-Item zur Selbsteinschätzung der Kompetenz im Verstehen der deutschen Sprache verteilten sich in beiden Erhebungswellen auf die drei oberen Antwortkategorien (s. *Abbildung 13* und *Abbildung 14* in Abschnitt 4.2.1). Die oberste Antwortkategorie wählten 27% der Teilnehmenden in der ersten Erhebungswelle und 48% der Teilnehmenden in der siebten Erhebungswelle, was auf einen deutlichen Deckeneffekt schließen lässt. Die Variation der objektiv gemessenen Kompetenzen streckt sich auf einen sehr viel größeren Wertebereich mit deutlich differenzierteren Kategorien und unterliegt keinem Deckeneffekt. Erwartungsgemäß war die Variation der Selbsteinschätzungen auf dem Standard-Item somit deutlich geringer als die der Ergebnisse der Deutschkompetenztests.

Bei den Schieberegler-Items und den Vergleich-Items haben die Teilnehmenden die vorgegebenen Skalen in ihrer Breite besser ausgenutzt und die Antworten verteilten sich auf mehr Kategorien als bei den Standard-Items (s. *Abbildung 15* und *Abbildung 16* in Abschnitt 4.2.1). Jedoch wählten die Teilnehmenden auch bei den Schieberegler-Items bevorzugt die höheren Kategorien, sodass sich der größte Teil der Antworten auf nur sechs Antwortkategorien verteilte und auch hier ein zumindest kleiner Deckeneffekt auftrat. Auch bei den Vergleich-Items konnten die vier überwiegend gewählten Kategorien die Variation der objektiv gemessenen Werte kaum angemessen wiedergeben. Fast die Hälfte der Teilnehmenden wählten dieselbe Kategorie, sodass die Variation deren tatsächlicher Kompetenz nicht abgebildet wurde.

Die Summenwerte der Can-Do-Statements verteilen sich auf acht Antwortkategorien. Auch hier zeigt sich, dass der größte Teil der Teilnehmenden min. vier oder fünf Can-Do-Statements

ausgewählt hat. Weiterhin zeigt sich in beiden Erhebungswellen ein Deckeneffet, der in der siebten Erhebungswelle besonders deutlich ausgeprägt ist, da hier 49% der Teilnehmenden alle der sieben in der Skala enthaltenen Can-Do-Statements zum Verstehen und Sprechen ausgewählt haben. Das heißt, auch mit dem Summenwert der Can-Do-Statements kann die Variation der sprachlichen Kompetenzen im Deutschen der jugendlichen Flüchtlinge nicht angemessen abgebildet werden.

Hypothese 1c wurde demnach für alle Arten von Items bestätigt: deren Variation ist deutlich geringer als die der objektiv gemessenen Kompetenzen.

4.5 Einflussfaktoren der Selbsteinschätzungen

Zur Prüfung der Hypothesen 2a – 2e, dass die Leistung in Mathematik, die Fähigkeit zum schlussfolgernden Denken, das Engagement beim Deutschlernen, die Teilnahme an einem Deutschkurs und die Teilnahme an einem Deutschtest die Selbsteinschätzungen der Deutschkompetenzen der jugendlichen Flüchtlinge unabhängig von deren objektiv gemessener Kompetenz beeinflussen, habe ich Strukturgleichungsmodelle angewendet. Um den Anteil der Selbsteinschätzung zu modellieren, der unabhängig von der objektiv gemessenen Kompetenz ist, habe ich zwei verschiedene Varianten der Modellierung gewählt, welche gegenseitig zur Prüfung der Robustheit der Ergebnisse dienen. Im Folgenden beschreibe ich zunächst die beiden Modellierungsvarianten, welche in *Abbildung 28* und *Abbildung 29* grafisch dargestellt sind.

In beiden Modellierungsvarianten wurden die verschiedenen Faktoren, deren Einfluss auf die Selbsteinschätzungen vorhergesagt wurde, identisch modelliert. Die Leistung in Mathematik, die Fähigkeit zum schlussfolgernden Denken und das Engagement zum Deutschlernen wurden latent modelliert. Die jeweiligen Indikatoren sind in den genannten Abbildungen ersichtlich. Für das Engagement zum Deutschlernen diente ausschließlich der Summenwert zur Nutzung von Möglichkeiten zum Deutschlernen als Indikator. Da bei nur einem Indikator der Messfehler nicht geschätzt werden, jedoch auch keine Messfehlerfreiheit angenommen werden konnte, wurde die Fehlervarianz des Indikators auf die Funktion dessen Reliabilität ($(1 - \alpha) \cdot SD^2 = (1 - .684) \cdot 2.532 = 0.800$) festgelegt, da mit 1 − Reliabilität der fehlerbedingte Anteil an der gesamten Varianz des Indikators geschätzt wird (Kline, 2016). Die Teilnahme an einem Deutschkurs und die Teilnahme an einem Deutschtest gingen als manifeste Variablen in das Modell ein. Die beiden Modellierungsvarianten unterscheiden sich darin, wie die Selbsteinschätzung so modelliert wurde, dass der Einfluss der tatsächlichen bzw. gemessenen Deutschkompetenz kontrolliert werden konnte.

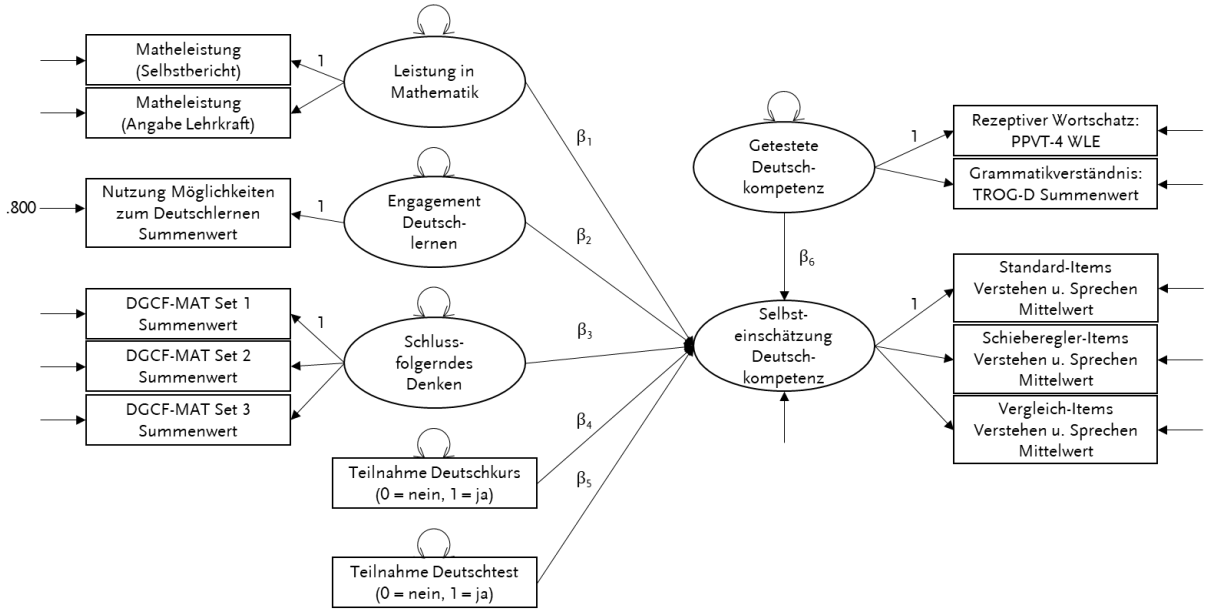
In der ersten Modellierungsvariante wurde die getestete Deutschkompetenz mit dem WLE des PPVT-4 und dem Summenwert der 47 Items des TROG-D als Indikatoren latent modelliert. Für

die latente selbsteingeschätzte Deutschkompetenz wurden jeweils der Mittelwert aus dem Item zum Verstehen und dem Item zum Sprechen der Standard-Items, der Schieberegler-Items und der Vergleich-Items als Indikatoren herangezogen[4]. Im Strukturmodell der ersten Modellierungsvariante wurden zur Prüfung der Hypothesen die Selbsteinschätzung der Deutschkompetenz auf die verschiedenen Faktoren sowie auf die getestete Deutschkompetenz regressiert, um für die getestete Deutschkompetenz zu kontrollieren. Die verschiedenen Faktoren durften untereinander sowie mit der getesteten Deutschkompetenz frei kovariieren. Die gerichteten Pfade der verschiedenen Faktoren auf die Selbsteinschätzung der Deutschkompetenz geben somit an, inwiefern die verschiedenen Faktoren die Selbsteinschätzungen unter Kontrolle der objektiv gemessenen Deutschkompetenzen in positive oder negative Richtung beeinflusst haben.

In der zweiten Modellierungsvariante waren die Testwerte und die Selbsteinschätzungen Indikatoren einer gemeinsamen latenten Variable für die Deutschkompetenz. Es wurden sowohl der WLE des PPVT-4 und der Summenwert der 47 Items des TROG-D als auch die drei Mittelwerte aus den Verstehen- und Sprechen-Items für die jeweiligen Selbsteinschätzungsarten als Indikatoren der Deutschkompetenz herangezogen. Für die drei manifesten Selbsteinschätzungsvariablen wurde darüber hinaus ein von der Deutschkompetenz unabhängiger latenter Methodenfaktor modelliert, dieser kennzeichnete die Abweichung der wahren Selbsteinschätzung von der wahren Deutschkompetenz (s. Eid et al., 2015). Zur Prüfung der Hypothesen wurde der Methodenfaktor der Selbsteinschätzung auf die verschiedenen Faktoren regressiert. Wie auch in der ersten Modellierungsvariante durften die verschiedenen Faktoren untereinander sowie mit der latenten Deutschkompetenzvariablen frei kovariieren. Die gerichteten Pfade der verschiedenen Faktoren auf den Methodenfaktor der Selbsteinschätzung der Deutschkompetenz geben in dieser Modellierungsvariante an, inwiefern die verschiedenen Faktoren die Abweichung der wahren Selbsteinschätzung von der wahren Deutschkompetenz in positive oder negative Richtung beeinflusst haben.

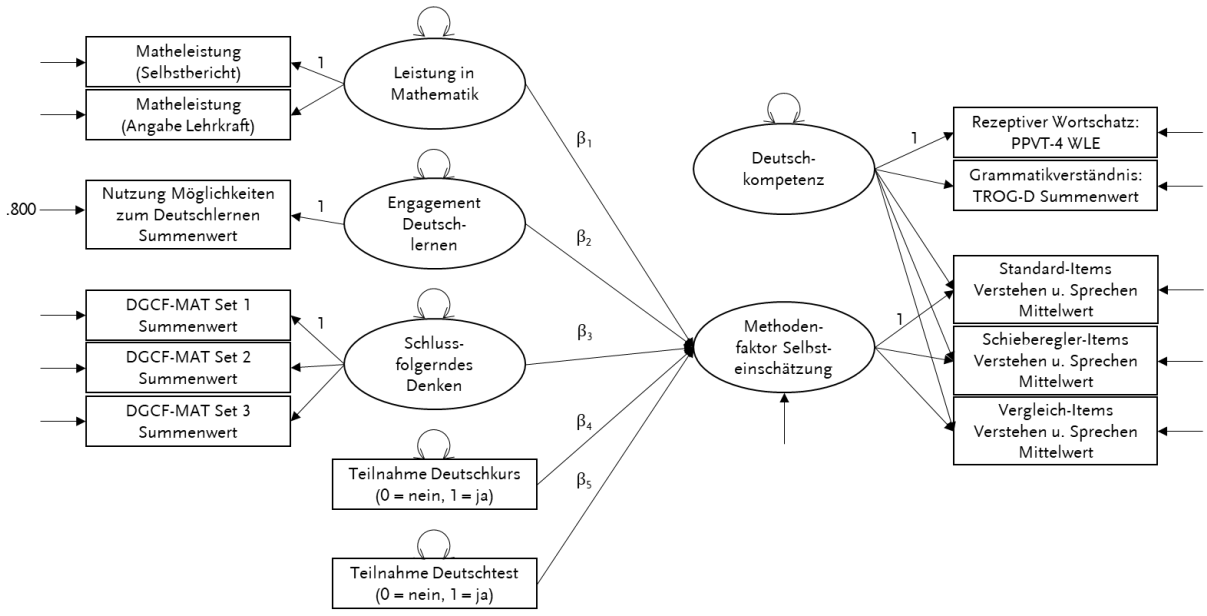
[4] Die Can-Do-Statements wurden hier nicht einbezogen, weil sie sich in ihrer Art der Selbsteinschätzung von den anderen Selbsteinschätzungsitems unterscheiden und sie sich somit nicht in den latenten Selbsteinschätzungsfaktor integrieren lassen. Zudem könnten bei diesen Items auch andere Einflussfaktoren eine Rolle spielen, als bei den anderen Items.

Abbildung 28. Erste Modellierungsvariante des Strukturgleichungsmodells zu den Einflussfaktoren von Selbsteinschätzungen



Anmerkungen. In dieser Modellierungsvariante wird der Einfluss der objektiven Deutschkompetenz auf die Selbsteinschätzung durch einen direkten Pfad der getesteten Deutschkompetenz auf die latente Selbsteinschätzungsvariable kontrolliert. Korrelationen zwischen allen latenten Variablen und den Variablen zur Teilnahme an einem Deutschkurs und an einem Deutschtest waren zugelassen, wurden wegen der Übersichtlichkeit jedoch nicht ins Modell eingezeichnet.

Abbildung 29. Zweite Modellierungsvariante des Strukturgleichungsmodells zu den Einflussfaktoren von Selbsteinschätzungen



Anmerkungen. In dieser Modellierungsvariante sind sowohl die Kompetenztests als auch die Selbsteinschätzungen Indikatoren der latenten Variablen zur Deutschkompetenz. Darüber hinaus wird der von der Deutschkompetenz unabhängige Methodenfaktor der Selbsteinschätzung modelliert. Korrelationen

zwischen allen latenten Variablen und den Variablen zur Teilnahme an einem Deutschkurs und an einem Deutschtest waren zugelassen, wurden wegen der Übersichtlichkeit jedoch nicht ins Modell eingezeichnet.

Die Strukturgleichungsmodelle wurden mithilfe des lavaan-Pakets Version 0.6-12 in R Version 4.0.3 (R Core Team, 2020) anhand der Rohdaten geschätzt. Es wurde der ML-Schätzer gewählt. Die verwendeten Daten umfassten die $N = 1\,877$ Fälle der Analysestichprobe 1. Um fehlende Werte in die Analysen miteinzubeziehen, wurde das Full-Information-Maximum-Likelihood-Schätzverfahren (FIML-Schätzverfahren) angewendet.[5]

Es wurde ein zweistufiges Vorgehen gewählt, bei dem jeweils zuerst das Messmodell geschätzt wurde, das die beiden manifesten Variablen ausschloss und freie Kovariationen zwischen allen latenten Variablen zuließ, mit Ausnahme der Kovariation zwischen der latenten Variablen für die Deutschkompetenz und dem latenten Methodenfaktor der Selbsteinschätzung in der zweiten Modellierungsvariante, welche auf null fixiert wurde. Im zweiten Schritt wurde jeweils das gesamte Modell inkl. dem beschriebenen Strukturmodell geschätzt und somit die Hypothesen geprüft. Im Folgenden werden die Ergebnisse in der Reihenfolge des beschriebenen zweistufigen Vorgehens berichtet. Das Vorgehen sowie der Bericht der Ergebnisse orientieren sich insbesondere an den Ausführungen von Kline (2016).

Ergebnisse für ausgewählte Statistiken zur Modellgüte für die beiden Messmodelle finden sich in *Tabelle 15*. Die Modelle scheinen die Daten gut zu repräsentieren. Unter den Residuen der verschiedenen Indikatoren treten jedoch Korrelationen in problematischer Höhe auf (s. *Tabelle D 1* und *Tabelle D 2* in Anhang D). Im Messmodell der ersten und zweiten Modellierungsvariante liegen jeweils vier Werte knapp über |.10|. Betroffen ist jeweils die Lehrerangabe zur Mathenote, deren Residuum mit den Residuen der Standard-Items und der Vergleich-Items der Selbsteinschätzung sowie mit den Residuen des zweiten und dritten Sets des DGCF-MAT korreliert. Weiterhin korrelieren die Residuen der Lehrerangabe der Matheleistung und des TROG-D im Messmodell der Modellierungsvariante 1 mit $r = .16$ und im Messmodell der Modellierungsvariante 2 ebenfalls mit $r = .16$, die Residuen der Lehrerangabe der Mathenote und des PPVT-4 korrelieren mit $r = .24$ und $r = .23$ in der ersten bzw. zweiten Modellierungsvariante des Messmodells und die Residuen der mit dem Schieberegler und den Vergleichs-Items gemessenen Selbsteinschätzung korrelieren mit $r = -.28$ und $r = -.31$ in den jeweiligen Modellierungsvarianten. Insbesondere die Lehrerangabe der Mathenote scheint hier problematisch zu sein, was u.a. mit dem hohen Anteil fehlender Werte auf dieser Variable zusammenhängen könnte. Eine fundierte theoretische Begründung für die Korrelationen der Residuen der Lehrerangabe der Mathenote wurde jedoch nicht gefunden, weshalb auch für weitere Analysen keine Korrelation zwischen den betroffenen Residuen im Modell

[5] Die R-Skripte, die die Modellspezifikationen in lavaan enthalten, werden auf Anfrage zur Verfügung gestellt.

zugelassen wurde. Auch für die Korrelation der Residuen der Schieberegler- und Vergleich-Items wurde keine theoretische Begründung gefunden, u.a. da die Standard-Items nicht betroffen waren und der Methodenfaktor der Selbsteinschätzungen in der zweiten Variante modelliert wurde. Deshalb wurde auch für diese Residuen keine Korrelation in den Modellen zugelassen.

Tabelle 15. Ausgewählte Indizes der Modellgüte für die Messmodelle und Gesamtmodelle zu den Einflussfaktoren der Selbsteinschätzungen (1. und 2. Modellierungsvariante)

Modell	χ^2	*df*	*p*	CFI	RMSEA (90% KI)	SRMR
Messmodelle						
1. Modellierungsvariante	72.235	35	<.001	.989	.024 [.016, .032]	.057
2. Modellierungsvariante	66.371	33	.001	.990	.023 [.015, .031]	.058
Gesamtmodelle						
1. Modellierungsvariante	112.349	49	<.001	.984	.026 [.020, .033]	.049
2. Modellierungsvariante	107.640	47	<.001	.985	.026 [.020, .033]	.050

In *Tabelle 16* sind für die Faktorladungen und für die Fehlervarianzen der Indikatoren die unstandardisierten Schätzer, deren Standardfehler sowie die standardisierten Schätzer im Messmodell der ersten Modellierungsvariante gelistet. In diesem Messmodell sind die standardisierten Faktorladungen überwiegend hoch, die Ladung des Indikators des ersten Sets des DGCF-MAT auf den Faktor schlussfolgerndes Denken stellt dabei mit einem niedrigeren Wert von .55 eine Ausnahme dar. Die Aufklärung der Varianz der verschiedenen Indikatoren durch die Faktoren liegt im Bereich von $R^2 = .30$ bis $R^2 = .72$.

Tabelle 16. ML-Schätzer der Faktorladungen und Fehlervarianzen im Messmodell zu den Einflussfaktoren der Selbsteinschätzungen (1. Modellierungsvariante)

	Faktorladungen			Fehlervarianzen		
	Unstandardisiert		Standardisiert	Unstandardisiert		Standardisiert
Indikator	**Schätzer**	**SE**	**Schätzer**	**Schätzer**	**SE**	**Schätzer**
Getestete Deutschkompetenz						
Rezeptiver Wortschatz: PPVT-4	1.00	-	.85	0.89	0.10	.28
Grammatikverständnis: TROG-D	3.34	0.15	.85	9.63	1.08	.28
Selbsteinschätzung						
Standard-Items	1.00	-	.78	0.13	0.01	.40

	Faktorladungen			Fehlervarianzen		
	Unstandardisiert		Standardisiert	Unstandardisiert		Standardisiert
Indikator	Schätzer	SE	Schätzer	Schätzer	SE	Schätzer
Schieberegler-Items	2.90	0.17	.84	0.70	0.10	.29
Vergleich-Items	1.16	0.08	.71	0.27	0.02	.50
Leistung in Mathematik						
Selbstberichtete Mathenote	1.00	-	.77	0.58	0.20	.41
Mathenote Lehrerangabe	1.00	0.25	.68	0.96	0.22	.54
Engagement beim Deutschlernen						
Nutzung Möglichkeiten zum Deutschlernen	1.00	-	.83	0.80	-	.32
Schlussfolgerndes Denken						
DGCF-MAT Set 1	1.00	-	.55	0.73	0.03	.70
DGCF-MAT Set 2	1.67	0.10	.74	0.73	0.05	.46
DGCF-MAT Set 3	1.51	0.09	.70	0.72	0.04	.50

In *Tabelle 17* sind für die Faktorladungen und für die Fehlervarianzen der Indikatoren die unstandardisierten Schätzer, deren Standardfehler sowie die standardisierten Schätzer im Messmodell der zweiten Modellierungsvariante gelistet. In diesem Messmodell laden die beiden Kompetenztests sehr hoch (jeweils .85) auf dem Faktor Deutschkompetenz, die Selbsteinschätzungen laden auf diesem Faktor niedrig mit Werten zwischen .35 und .37. Auf dem Methodenfaktor der Selbsteinschätzung laden die drei Selbsteinschätzungen deutlich höher mit Schätzwerten zwischen .62 und .81. Alle weiteren Indikatoren laden hoch auf den jeweiligen Faktoren, wobei auch in diesem Modell das erste Set des DGCF-MAT eine Ausnahme darstellt, mit einem Schätzwert von .55. Die Aufklärung der Varianz der verschiedenen Indikatoren durch die Faktoren liegt im Bereich von $R^2 = .30$ bis $R^2 = .77$.

Tabelle 17. ML-Schätzer der Faktorladungen und Fehlervarianzen im Messmodell zu den Einflussfaktoren der Selbsteinschätzungen (2. Modellierungsvariante)

	Faktorladungen			Fehlervarianzen		
	Unstandardisiert		Standardisiert	Unstandardisiert		Standardisiert
Indikator	Schätzer	*SE*	Schätzer	Schätzer	*SE*	Schätzer
Deutschkompetenz						
Rezeptiver Wortschatz: PPVT-4	1.00	-	.85	0.88	0.10	.28
Grammatikverständnis: TROG-D	3.32	0.15	.85	9.79	1.08	.28
Standard-Items	0.14	0.01	.36	0.14	0.02	.43

	Faktorladungen			Fehlervarianzen		
	Unstandardisiert		Standardisiert	Unstandardisiert		Standardisiert
Indikator	Schätzer	*SE*	Schätzer	Schätzer	*SE*	Schätzer
Schieberegler-Items	0.36	0.03	.35	0.54	0.21	.23
Vergleich-Items	0.18	0.02	.37	0.26	0.03	.49
Methodenfaktor Selbsteinschätzung						
Standard-Items	1.00	-	.66	0.14	0.02	.43
Schieberegler-Items	3.27	0.43	.81	0.54	0.21	.23
Vergleich-Items	1.18	0.16	.62	0.26	0.03	.49
Leistung in Mathematik						
Selbstberichtete Mathenote	1.00	-	.76	0.60	0.20	.42
Mathenote Lehrerangabe	1.02	0.26	.69	0.94	0.22	.53
Engagement beim Deutschlernen						
Nutzung Möglichkeiten zum Deutschlernen	1.00	-	.83	0.80	-	.32
Schlussfolgerndes Denken						
DGCF-MAT Set 1	1.00	-	.55	0.73	0.03	.70
DGCF-MAT Set 2	1.67	0.10	.74	0.73	0.05	.46
DGCF-MAT Set 3	1.52	0.09	.70	0.72	0.04	.50

Die Schätzer der Faktorvarianzen und -kovarianzen für beide Messmodelle sind in *Tabelle 18* gelistet. Die Schätzer der Faktorkorrelationen liegen für das Messmodell der ersten Modellierungsvariante in einem moderaten Bereich von .10 bis .45. Für das Messmodell 2 liegen die Schätzer der Faktorkorrelationen ebenfalls in einem moderaten Bereich von −.10 bis .43. Dies spricht dafür, dass die Faktoren in beiden Modellen diskriminant valide sind.

Tabelle 18. ML-Schätzer der Faktorvarianzen und Faktorkovarianzen in den Messmodellen zu den Einflussfaktoren der Selbsteinschätzungen (1. und 2. Modellierungsvariante)

	Messmodell 1. Modellierungsvariante			Messmodell 2. Modellierungsvariante		
	Unstandardisiert		Standardisiert	Unstandardisiert		Standardisiert
Faktor(en)	Schätzer	*SE*	Schätzer	Schätzer	*SE*	Schätzer
			Faktorvarianzen			
(getestete) Deutschkompetenz	2.24	0.14	1.00	2.25	0.14	1.00
(Methodenfaktor) Selbsteinschätzung	0.20	0.01	1.00	0.15	0.02	1.00
Leistung in Mathematik	0.83	0.20	1.00	0.81	0.20	1.00
Engagement beim Deutschlernen	1.73	0.08	1.00	1.73	0.08	1.00

	Messmodell 1. Modellierungsvariante			Messmodell 2. Modellierungsvariante		
	Unstandardisiert		Standardisiert	Unstandardisiert		Standardisiert
Faktor(en)	**Schätzer**	***SE***	**Schätzer**	**Schätzer**	***SE***	**Schätzer**
Schlussfolgerndes Denken	0.31	0.03	1.00	0.31	0.03	1.00
			Faktorkovarianzen			
(getestete) Deutschkompetenz ↔ (Methodenfaktor) Selbsteinschätzung	0.30	0.02	.45	0.00	-	.00
(getestete) Deutschkompetenz ↔ Leistung in Mathematik	0.30	0.05	.22	0.30	0.05	.22
(getestete) Deutschkompetenz ↔ Engagement beim Deutschlernen	0.35	0.06	.18	0.35	0.06	.18
(getestete) Deutschkompetenz ↔ Schlussfolgerndes Denken	0.36	0.03	.43	0.36	0.03	.43
(Methodenfaktor) Selbsteinschätzung ↔ Leistung in Mathematik	0.13	0.02	.32	0.09	0.02	.25
(Methodenfaktor) Selbsteinschätzung ↔ Engagement beim Deutschlernen	0.12	0.02	.21	0.07	0.02	.15
(Methodenfaktor) Selbsteinschätzung ↔ Schlussfolgerndes Denken	0.03	0.01	.10	-0.02	0.01	−.10
Leistung in Mathematik ↔ Engagement beim Deutschlernen	0.29	0.05	.24	0.29	0.05	.25
Leistung in Mathematik ↔ Schlussfolgerndes Denken	0.17	0.02	.33	0.17	0.02	.33
Engagement beim Deutschlernen ↔ Schlussfolgerndes Denken	0.14	0.03	.20	0.14	0.03	.20

Als nächstes werden die Ergebnisse des zweiten Schritts der Analysen beschrieben, in dem die Gesamtmodelle analysiert wurden, welche neben den Messmodellen die Teilnahme an Deutschunterricht und einem Deutschtest als manifeste Variablen sowie die für die Hypothesentestung relevanten Spezifikationen des Strukturmodells enthielten. Die Angaben zur Modellgüte in *Tabelle 15* sprechen für eine gute Passung zwischen Modell und Daten. Problematische Korrelationen der Residuen finden sich in den Gesamtmodellen in sehr ähnlichem Ausmaß wie in den Messmodellen und werden hier nicht im Detail ausgeführt, sind aber in *Tabelle D 3* für das Gesamtmodell der ersten Modellierungsvariante und in *Tabelle D 4* für das Gesamtmodell der zweiten Modellierungsvariante in Anhang D enthalten. Zwischen den Residuen der hinzugekommenen Variablen zur Teilnahme an Deutschunterricht und zur Teilnahme an einem Deutschtest überstieg nur die Korrelation zwischen der Teilnahme an Deutschunterricht und dem Summenwert des TROG-D den Grenzwert von |.10| knapp.

Das gesamte Modell inklusive der Parameterschätzer wird für die erste Modellierungsvariante in *Abbildung 30* und für die zweite Modellierungsvariante in *Abbildung 31* dargestellt. Die

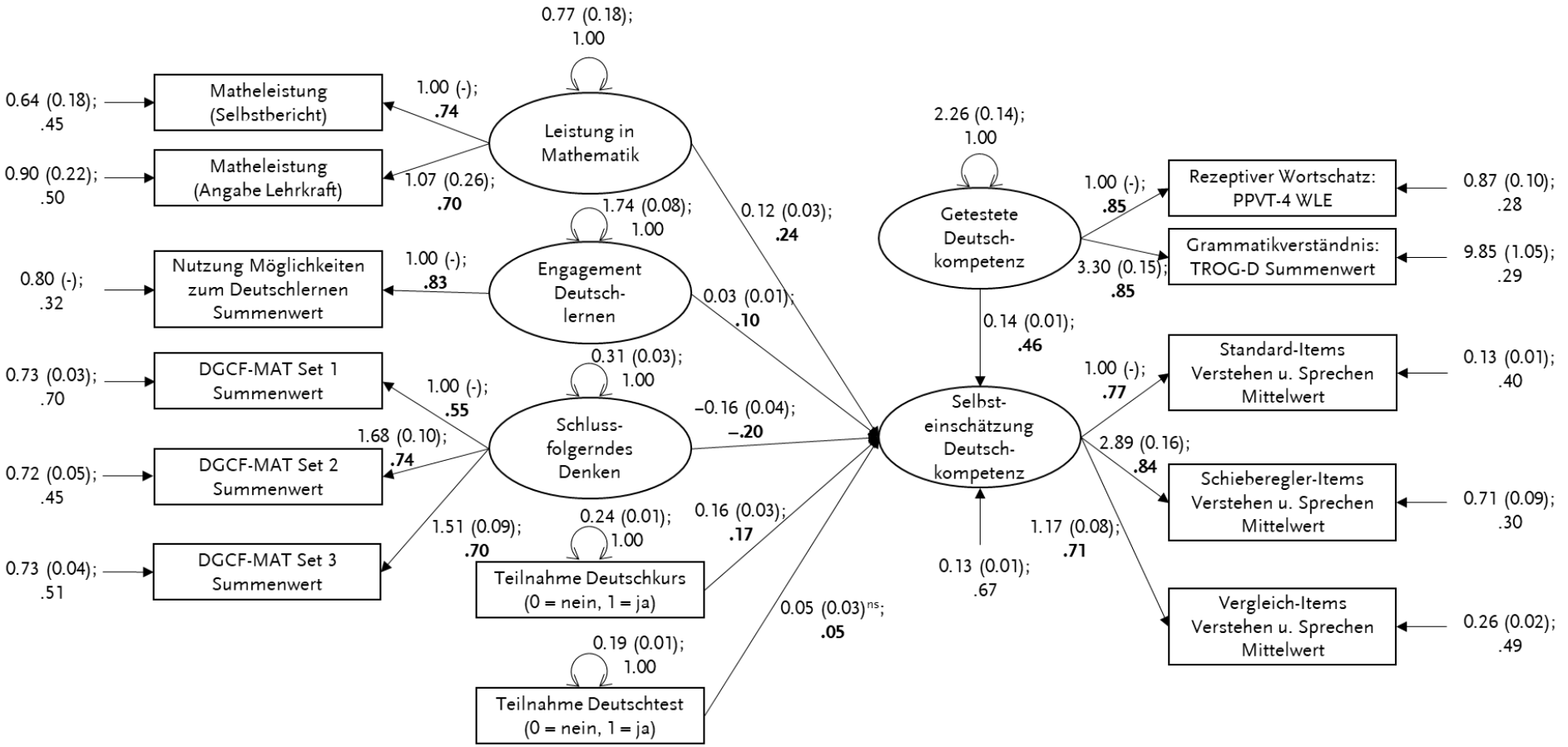
geschätzten Ladungen der verschiedenen Indikatoren auf den latenten Faktoren ähneln denen der Messmodelle stark, weshalb hier nicht erneut auf diese eingegangen wird. Die Aufklärung der Varianz der Indikatoren durch die latenten Faktoren liegt im Bereich von $R^2 = .30$ bis $R^2 = .72$ im Gesamtmodell der ersten Modellierungsvariante und im Bereich von $R^2 = .30$ bis $R^2 = .73$ im Gesamtmodell der zweiten Modellierungsvariante.

Für die Hypothesenprüfung sind die direkten Pfade der verschiedenen Einflussfaktoren auf die latente Variable für die Selbsteinschätzung in der ersten Modellierungsvariante bzw. den Methodenfaktor der Selbsteinschätzung in der zweiten Modellierungsvariante relevant. Zuerst werden die Ergebnisse der Hypothesenprüfung anhand des Modells mit der ersten Modellierungsvariante berichtet. Die Leistung in Mathematik hatte einen signifikant positiven Effekt auf die Selbsteinschätzung der Deutschkompetenz ($\hat{\beta}_1 = 0.12$, $z = 3.65$, $p < .001$, $\hat{\beta}_1^{standardisiert} = .24$), was der Hypothese 2a widerspricht, nach welcher aufgrund von internalen Referenzrahmen ein negativer Effekt der Matheleistung auf die Selbsteinschätzung der Deutschkompetenz vorhergesagt wurde. Das Engagement beim Deutschlernen hatte einen signifikant positiven Effekt auf die Selbsteinschätzung der Deutschkompetenz ($\hat{\beta}_2 = 0.03$, $z = 2.86$, $p = .004$, $\hat{\beta}_2^{standardisiert} = .10$), was die Hypothese 2b bestätigt. Die Effektstärke war jedoch gering. Bei vergleichbarer getesteter Deutschkompetenz schätzten demnach diejenigen jugendlichen Flüchtlinge, die ein höheres Engagement beim Deutschlernen angaben, ihre Deutschkompetenz etwas höher ein als jugendliche Flüchtlinge, die weniger Engagement beim Deutschlernen angaben. Die Fähigkeit zum schlussfolgernden Denken hatte einen signifikant negativen Effekt auf die Selbsteinschätzung der Deutschkompetenz ($\hat{\beta}_3 = -0.16$, $z = -4.35$, $p < .001$, $\hat{\beta}_3^{standardisiert} = -.20$), was die Hypothese 2c bestätigt. Bei vergleichbarer getesteter Deutschkompetenz schätzten demnach Teilnehmende mit höher ausgeprägter Fähigkeit zum schlussfolgernden Denken ihre Deutschkompetenz niedriger ein als Teilnehmende mit niedriger ausgeprägter Fähigkeit zum schlussfolgernden Denken. Bei gleicher getesteter Deutschkompetenz schätzten diejenigen, die an einem Deutschkurs teilgenommen hatten, ihre Deutschkompetenz höher ein als diejenigen, die an keinem Deutschkurs teilgenommen hatten ($\hat{\beta}_4 = 0.16$, $z = 5.80$, $p < .001$, $\hat{\beta}_4^{standardisiert} = .17$). Dieser Befund widerspricht der Hypothese 2d, die einen gegenteiligen Effekt vorhergesagt hatte. Die Teilnahme an einem Deutschtest zusätzlich zur Teilnahme an einem Deutschkurs hatte keinen signifikanten Einfluss auf die Selbsteinschätzung der Deutschkompetenz ($\hat{\beta}_5 = 0.05$, $z = 1.63$, $p = .10$, $\hat{\beta}_5^{standardisiert} = .05$). Hypothese 2e wurde somit nicht bestätigt, wonach die zusätzliche Teilnahme an einem Deutschtest einen negativen Effekt auf die Selbsteinschätzung der Deutschkompetenz haben sollte.

Die Ergebnisse zur Hypothesenprüfung anhand des Modells mit der zweiten Modellierungsvariante decken sich mit den berichteten Ergebnissen zur Hypothesenprüfung anhand des Modells mit der ersten Modellierungsvariante. Die Leistung in Mathematik hatte einen signifikant positiven

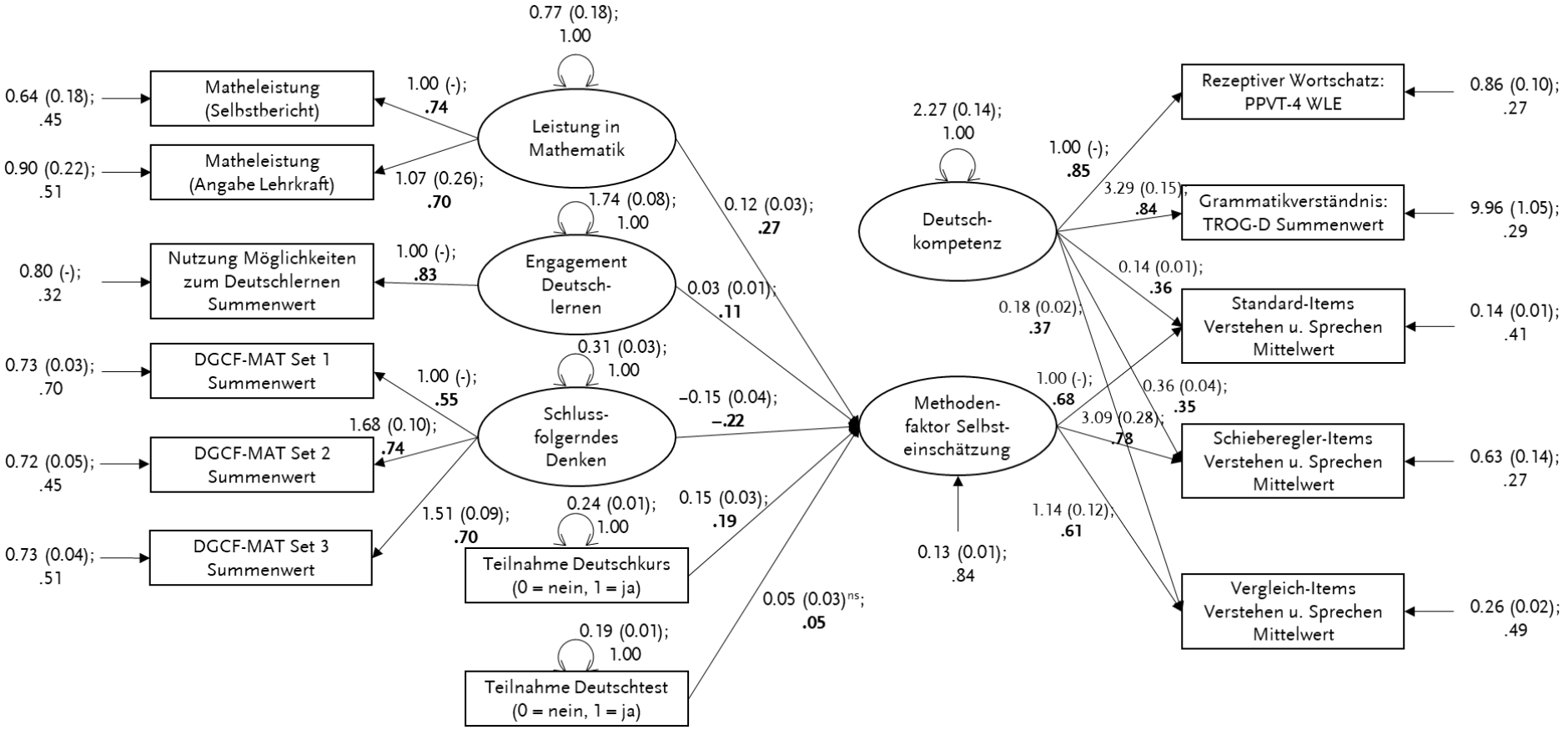
Effekt auf den Methodenfaktor der Selbsteinschätzung ($\hat{\beta}_1 = 0.12$, $z = 3.65$, $p < .001$, $\hat{\beta}_1^{standardisiert} = .27$), was Hypothese 2a widerspricht. Das Engagement beim Deutschlernen hatte einen signifikant positiven Effekt auf den Methodenfaktor der Selbsteinschätzung ($\hat{\beta}_2 = 0.03$, $z = 2.83$, $p = .005$, $\hat{\beta}_2^{standardisiert} = .11$), was die Hypothese 2b bestätigt. Die Fähigkeit zum schlussfolgernden Denken hatte einen signifikant negativen Effekt auf den Methodenfaktor der Selbsteinschätzung ($\hat{\beta}_3 = -0.15$, $z = -4.26$, $p < .001$, $\hat{\beta}_3^{standardisiert} = -.22$), was die Hypothese 2c bestätigt. Hypothese 2d wurde auch in dieser Modellierungsvariante widersprochen, da die Teilnahme an einem Deutschkurs einen signifikant positiven Einfluss auf den Methodenfaktor der Selbsteinschätzung hatte ($\hat{\beta}_4 = 0.15$, $z = 5.62$, $p < .001$, $\hat{\beta}_4^{standardisiert} = .19$). Ebenso wurde Hypothese 2e auch hier nicht bestätigt, da die Teilnahme an einem Deutschtest zusätzlich zur Teilnahme an einem Deutschkurs keinen signifikanten Einfluss auf den Methodenfaktor der Selbsteinschätzung hatte ($\hat{\beta}_5 = 0.05$, $z = 1.56$, $p = .12$, $\hat{\beta}_5^{standardisiert} = .05$). Zusammenfassend bestätigen die Ergebnisse beider Gesamtmodelle die Hypothesen 2b und 2c, wonach jugendliche Flüchtlinge, die ein höheres Engagement beim Deutschlernen zeigen sowie jugendliche Flüchtlinge mit niedrigeren Fähigkeiten zum schlussfolgernden Denken ihre Deutschkompetenz höher einschätzen als Teilnehmende, die ein geringeres Engagement zeigen bzw. höhere Fähigkeiten zum schlussfolgernden Denken besitzen, jeweils unter Kontrolle der getesteten Deutschkompetenz. Die Hypothesen 2a, 2d und 2e zum Einfluss der Leistung in Mathematik, der Teilnahme an einem Deutschkurs und der zusätzlichen Teilnahme an einem Deutschtest auf die Selbsteinschätzung wurden nicht bestätigt.

Abbildung 30. Strukturgleichungsmodell zu den Einflussfaktoren der Selbsteinschätzungen (1. Modellierungsvariante)



Anmerkungen. Es sind jeweils die unstandardisierten Parameterschätzer und deren Standardfehler in Klammern in der oberen Zeile und die standardisierten Parameterschätzer in der unteren Zeile angegeben und für die Faktorladungen und direkten Effekte hervorgehoben. Korrelationen zwischen allen latenten Variablen und den Variablen zur Teilnahme an einem Deutschkurs und an einem Deutschtest waren zugelassen, wurden wegen der Übersichtlichkeit jedoch nicht im Modell angegeben. Für alle unstandardisierten Schätzer ist $p < .01$, außer diese sind mit ns (= nicht signifikant) markiert.

Abbildung 31. Strukturgleichungsmodell zu den Einflussfaktoren der Selbsteinschätzungen (2. Modellierungsvariante)



Anmerkungen. Es sind jeweils die unstandardisierten Parameterschätzer und deren Standardfehler in Klammern in der oberen Zeile und die standardisierten Parameterschätzer in der unteren Zeile angegeben und für die Faktorladungen und direkten Effekte hervorgehoben. Korrelationen zwischen allen latenten Variablen und den Variablen zur Teilnahme an einem Deutschkurs und an einem Deutschtest waren zugelassen, wurden wegen der Übersichtlichkeit jedoch nicht im Modell angegeben. Für alle unstandardisierten Schätzer ist $p < .01$, außer diese sind mit ns (= nicht signifikant) markiert.

4.6 Vergleich verschiedener Arten von Selbsteinschätzungsitems

In Kapitel 2.3 wurden Hypothesen zu den verschiedenen Arten von Selbsteinschätzungsitems aufgestellt. Die Ergebnisse der Hypothesenprüfung werden im Folgenden berichtet. Für den Vergleich von Korrelationen der verschiedenen Arten von Selbsteinschätzungsitems mit den getesteten Deutschkompetenzen wurden jeweils nur die Fälle herangezogen, für die Werte zu allen für den Vergleich relevanten Variablen vorlagen, alle anderen Fälle wurden ausgeschlossen. Da die Schieberegler-Items und die Vergleich-Items jeweils nur von einer Hälfte der Stichprobe bearbeitet wurden, sind die Stichproben für diese Vergleiche entsprechend kleiner.[6] Gegebenenfalls habe ich zur Prüfung der Signifikanz des Unterschieds zwischen zwei abhängigen Korrelationen einen gerichteten t-Test mithilfe der r.test-Funktion des psych-Pakets Version 2.2.5 (Revelle, 2022) in R Version 4.0.3 (R Core Team, 2020) durchgeführt.

4.6.1 Schieberegler-Items

In Hypothese 3a wurde vorhergesagt, dass das Schieberegler-Item weniger schief sei als das Standard-Item zur Selbsteinschätzung der Deutschkompetenz. Die Hypothese wird durch die Daten jedoch widerlegt, da entsprechend den Werten für die Schiefe der Items, welche in *Tabelle 8* in Abschnitt 4.2.1 enthalten sind, die Verteilung des Schieberegler-Items schiefer ist als die des Standard-Items ($Schiefe_{Schieberegler} = -0.85$, $Schiefe_{Standard} = -0.39$).

Da die höhere Anzahl an Antwortkategorien insbesondere im oberen Bereich der Skala eine bessere Diskrimination ermöglicht, wurde in Hypothese 3b vorhergesagt, dass die Schieberegler-Items stärker mit den objektiv gemessenen Deutschkompetenzen korrelieren als die Standard-Items. Zur Prüfung dieser Hypothese wurden die Korrelation des Standard-Items zum Verstehen und des Schieberegler-Items zum Verstehen mit dem kombinierten Deutschkompetenzwert miteinander verglichen, wobei nur die $n = 944$ Fälle in den Vergleich miteingingen, für die keiner der drei Werte fehlte. Auch diese Hypothese wurde anhand der Daten widerlegt, da die Korrelation zwischen dem Schieberegler-Item und dem kombinierten Deutschkompetenzscore mit $r = .31$ etwas niedriger war als die Korrelation zwischen dem Standard-Item und dem kombinierten Deutschkompetenzscore mit $r = .32$.

[6] In *Tabelle 13* und *Tabelle 14* in Abschnitt 4.4.1 und in *Tabelle C 1* und *Tabelle C 2* in Anhang C wurden die Korrelationen für alle Fälle berechnet, für die die beiden in die Korrelation eingehenden Werte vorlagen. Deshalb können sich die Fallzahlen und Korrelationen in dieser Tabelle von den hier berichteten Fallzahlen und Korrelationswerten unterscheiden.

4.6.2 Vergleich-Items

Für die Verteilung der Antworten auf das Vergleich-Item wurde in Hypothese 4a vorhergesagt, dass diese keinen Deckeneffekt zeige. Die oberste Kategorie des Vergleichs-Items wurde nur von 5% der Teilnehmenden gewählt, deshalb werte ich Hypothese 4a als bestätigt.

Die Korrelation des Vergleich-Items mit dem kombinierten Deutschkompetenzscore ($r = .30$) war etwas höher als die Korrelation des Standard-Items mit dem kombinierten Deutschkompetenzscore ($r = .29$) in der Stichprobe von $n = 922$ Teilnehmenden, für die alle relevanten Werte vorlagen. Ein gerichteter t-Test zur Prüfung der Signifikanz des Unterschieds zwischen den beiden abhängigen Korrelationen ergab jedoch, dass der Unterschied nicht signifikant ist ($t(921) = -0.58$, $p = .28$). Somit wurde Hypothese 4b nicht bestätigt.

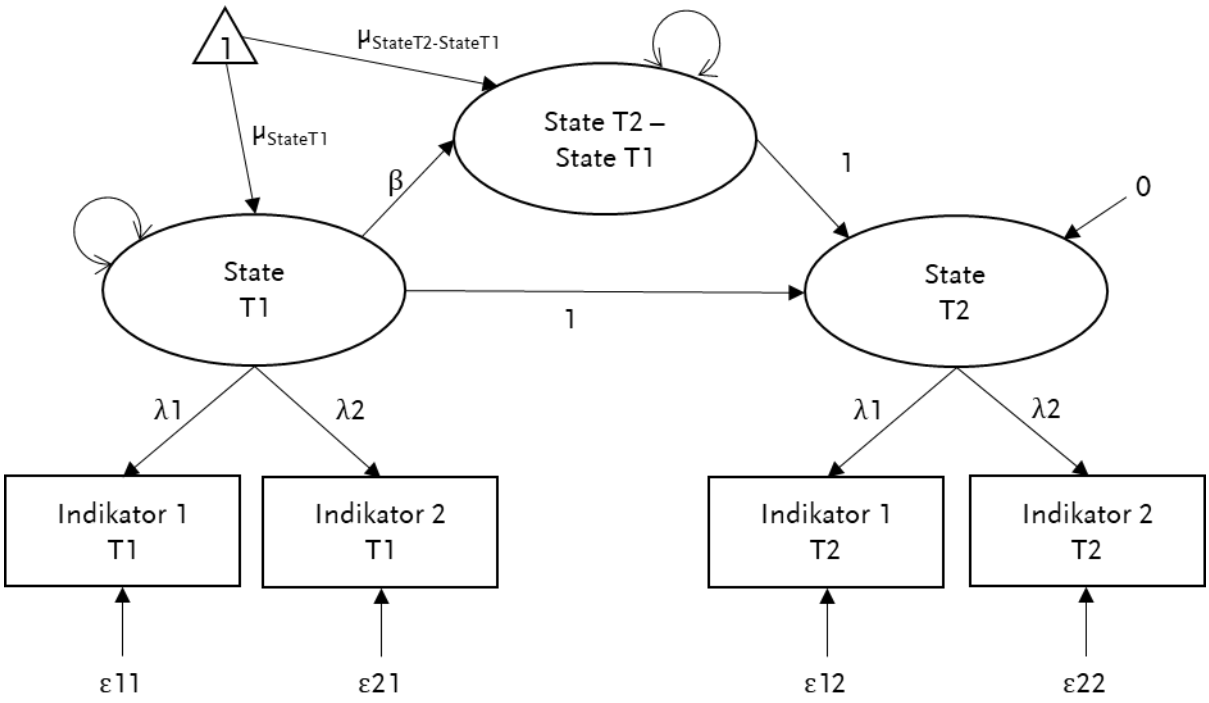
4.6.3 Can-Do-Statements

Da die Can-Do-Statements sowohl in der ersten als auch in der siebten Erhebungswelle Teil der Befragung waren, konnte die Hypothese 5a zum Vergleich der Korrelation der Can-Do-Statements zum Verstehen und Sprechen und der Standard-Items jeweils mit dem kombinierten Deutschkompetenzscore anhand der Daten beider Erhebungswellen getestet werden. In der ersten Erhebungswelle korrelierte der Summenwert der Can-Do-Statements zum Verstehen und Sprechen etwas schwächer mit dem kombinierten Deutschkompetenzscore ($r = .25$) als das Standard-Item zum Verstehen der deutschen Sprache ($r = .30$). In die Analyse gingen die Daten von $n = 1\,870$ Teilnehmenden ein. In der siebten Erhebungswelle korrelierte der Summenwert der Can-Do-Statements zum Verstehen und Sprechen ($r = .39$) etwas stärker mit dem kombinierten Deutschkompetenzscore als das Standard-Item ($r = .37$). In die Analyse gingen die Daten von $n = 777$ Teilnehmenden ein. Entsprechend dem gerichteten t-Test zur Prüfung der Signifikanz des Unterschieds zwischen den beiden abhängigen Korrelationen in der siebten Erhebungswelle war dieser Unterschied jedoch nicht signifikant ($t(776) = -0.63$, $p = .26$). Die Hypothese 5a konnte somit für beide Erhebungswellen nicht bestätigt werden.

Zur Prüfung der Hypothese 5b, dass die Can-Do-Statements Veränderungen der Kompetenz zwischen zwei Messzeitpunkten besser erfassen können als die Standard-Items, habe ich die Zusammenhänge zwischen den Veränderungen der beiden subjektiven Kompetenzmaße mit den Veränderungen der objektiven Kompetenzmaße als Kriterium miteinander verglichen. Da bei allen verwendeten Variablen davon auszugehen war, dass sie Messfehlern unterliegen, habe ich die Veränderungen mithilfe eines Latent-Change-Modells latent modelliert. Im Folgenden werden die Modellierung und die Analysen genauer beschrieben und die Ergebnisse berichtet.

In einem Latent-Change-Modell wird die Veränderung über die Zeit in Form von latenten Differenzvariablen wie in *Abbildung 32* allgemein dargestellt modelliert. Dazu wird der latente State-Faktor zum zweiten Messzeitpunkt in den Ausgangswert zum ersten Messzeitpunkt und die latente Differenz zerlegt (State T2 = State T1 + (State T2 – State T1)). Der latente State-Faktor zum zweiten Messzeitpunkt wird in dem Modell vollständig durch den latenten State-Faktor zum ersten Messzeitpunkt und die latente Differenzvariable determiniert, sodass es kein Residuum für den latenten State-Faktor zum zweiten Messzeitpunkt gibt. Latente Differenzvariablen können in einem Strukturgleichungsmodell wie andere Variablen mit weiteren Variablen in Beziehung gesetzt werden. Einem Latent-Change-Modell liegt die Annahme starker faktorieller Invarianz zugrunde, d.h., dass sich die Faktorladungen und die Intercepts der Indikatoren über die Messzeitpunkte nicht verändern. Würden sich die psychometrischen Eigenschaften der Indikatoren über die Zeit verändern, wären die latenten Variablen über die Messzeitpunkte nicht vergleichbar und die latenten Differenzvariablen ließen sich nicht sinnvoll interpretieren. Ob diese Annahme für die verschiedenen Kompetenzmaße haltbar ist, muss zuerst geprüft werden. Ist die Voraussetzung erfüllt, kann das Latent-Change-Modell mit den entsprechenden Restriktionen der starken faktoriellen Invarianz geschätzt werden (Geiser, 2011).

Abbildung 32. Allgemeines Latent-Change-Modell mit zwei beobachteten Variablen, die zu zwei Messzeitpunkten erhoben wurden



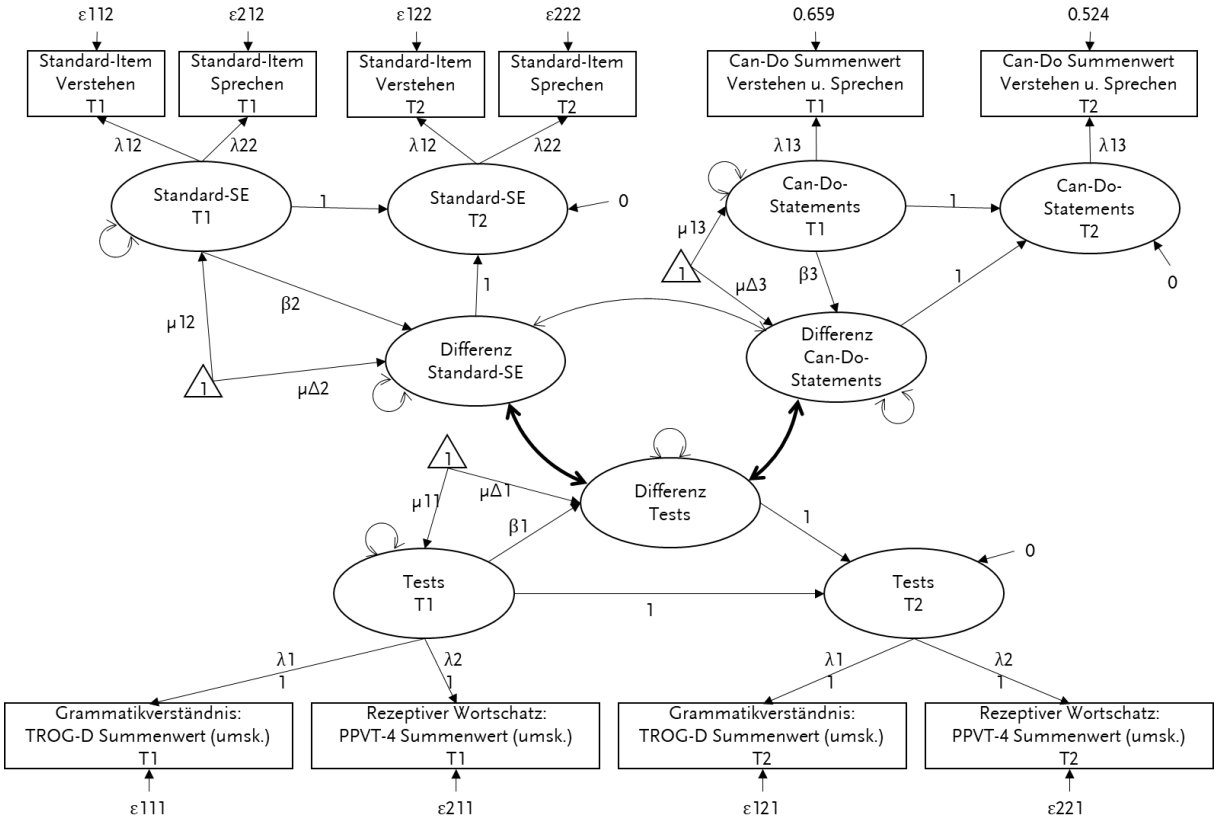
Anmerkungen. T1 = erster Messzeitpunkt; T2 = zweiter Messzeitpunkt; λi = zeitlich invariante Faktorladungsparameter; εim = Messfehlervariable; β = Regressionsparameter; μ = Mittelwert; i = Indikator; m = Messzeitpunkt. Adaptiert nach Geiser (2011) und Kievit et al. (2018).

Für die Analysen zur Hypothese 5b habe ich die latenten Differenzvariablen für jede der drei Arten von Kompetenzmaßen (Kompetenztests, Standard-Selbsteinschätzung, Can-Do-Statements) entsprechend den Ausführungen zum Latent-Change-Modell modelliert. Das Strukturgleichungsmodell ist in *Abbildung 33* dargestellt. Um wie auch bei den Can-Do-Statements die Kompetenzen im Verstehen und im Sprechen der deutschen Sprache einzubeziehen und zwei Indikatoren pro Messzeitpunkt zur Verfügung zu haben, damit die Messfehler geschätzt werden können, wurden als Indikatoren für die Standard-Items das Standard-Item zum Verstehen und das Standard-Item zum Sprechen der deutschen Sprache herangezogen. Für die Can-Do-Statements zum Verstehen und Sprechen wurde der Summenwert als alleiniger Indikator herangezogen. Um den Messfehler zu modellieren, wurde die Fehlervarianz des Indikators für jeden Messzeitpunkt auf die Funktion dessen Reliabilität ($(1 - \alpha) \cdot SD^2$) festgelegt (Messzeitpunkt 1: (1-0.765)*2.802 = 0.659; Messzeitpunkt 2: (1-0.809)*2.742 = 0.524; vgl. Vorgehen in Kapitel 4.5; Kline, 2016). Für die Kompetenztestung wurden umskalierte Summenwerte des PPVT-4[7] und umskalierte Summenwerte der 47 Items des TROG-D für die beiden Messzeitpunkte als Indikatoren herangezogen.[8] Das Strukturmodell enthielt alle drei Korrelation zwischen den latenten Differenzvariablen sowie die drei Korrelationen zwischen den latenten State-Faktoren der verschiedenen Kompetenzmaße zum ersten Messzeitpunkt. Weiterhin waren Regressionen der latenten Differenzvariablen auf die latenten State-Faktoren zum ersten Messzeitpunkt zugelassen. Für die Hypothesenprüfung waren die beiden Korrelationen der latenten Differenzvariablen der Selbsteinschätzungen mit der latenten Differenzvariable der Kompetenztests relevant, welche in *Abbildung 33* durch dickere Pfeile hervorgehoben sind.

[7] Der Summenwert des PPVT-4 wurde verwendet, weil keine (gelinkten) WLEs zum PPVT-4 zum zweiten Messzeitpunkt vorlagen und nur so die Skala zu beiden Messzeitpunkten identisch war und die Differenz modelliert werden konnte.

[8] Die Summenwerte wurden jeweils umskaliert, indem für jeden Test der Mittelwert des Summenwerts zum ersten Messzeitpunkt vom Summenwert subtrahiert wurde und das Ergebnis durch die Standardabweichung der Summenwerte zum ersten Messzeitpunkt geteilt wurde. Für den ersten Messzeitpunkt wurde also jeweils eine z-Standardisierung vorgenommen und für den zweiten Messzeitpunkt wurden für die Umskalierung der Mittelwert und die Standardabweichung zum ersten Messzeitpunkt verwendet, damit die Skalen zwischen den Messzeitpunkten vergleichbar blieben.

Abbildung 33. Strukturgleichungsmodell zur Veränderung der Kompetenzmaße über die Zeit



Anmerkungen. Im Modell werden die Veränderungen der verschiedenen Kompetenzmaße latent modelliert mithilfe von Latent-Change-Modellen (vgl. Geiser, 2011; Kievit et al., 2018) und zueinander in Beziehung gesetzt, sodass die hervorgehobenen Zusammenhänge verglichen werden können. Im Modell wurden weiterhin Kovarianzen zwischen den verschiedenen latenten State-Faktoren zum ersten Messzeitpunkt zugelassen sowie Regressionen der latenten Differenzvariablen auf die latenten State-Faktoren der anderen Kompetenzmaße zum ersten Messzeitpunkt. Diese sind aus Gründen der Übersichtlichkeit in der Abbildung nicht eingezeichnet. Standard-SE = Standard-Selbsteinschätzung; T1 = erster Messzeitpunkt; T2 = zweiter Messzeitpunkt; λik = zeitlich invariante Faktorladungsparameter; εimk = Messfehlervariable; βk = Regressionsparameter; μmk = Mittelwert; i = Indikator; m = Messzeitpunkt; k = Kompetenzmaß (1 = Tests, 2 = Standard-SE, 3 = Can-Do-Statements); umsk. = umskaliert (PPVT-4: (Summenwert − 89.42)/32.50; TROG-D: (Summenwert – 33.57)/5.71).

Für die Analysen wurde ein mehrstufiges Vorgehen gewählt. Zuerst wurde anhand des Messmodells, für das die latenten Differenzvariablen noch nicht modelliert wurden, die Annahme starker faktorieller Invarianz über die Zeit geprüft. Dazu wurden die Restriktionen schrittweise durch die Gleichsetzung von Parametern über die Zeit eingeführt und die Güte der Modelle jeweils miteinander verglichen, um festzustellen, ob auch das Modell mit den Restriktionen zur starken faktoriellen Invarianz zu den Daten passt. Zuerst wurde eine konfirmatorische Faktorenanalyse für ein Messmodell konfiguraler Invarianz mit den genannten Indikatoren und State-Faktoren für beide Messzeitpunkte ohne weitere Restriktionen geschätzt. Im nächsten Schritt wurde die Gleichheitsrestriktion für die Faktorladungen über die Zeit eingeführt. Damit wurde die schwache

faktorielle Invarianz des Messmodells geprüft. Nach dem Vergleich beider Modelle wurden zur Prüfung der starken faktoriellen Invarianz darüber hinaus die Intercepts der Indikatoren jeweils über die Zeit gleichgesetzt.[9] Dieses Modell wurde mit dem Modell schwacher faktorieller Invarianz verglichen. Das Vorgehen orientiert sich an den Ausführungen von Geiser (2011). Weiterhin wurden die Korrelationen zwischen Residuen und Parameterschätzer des Messmodells mit den Restriktionen zur starken faktoriellen Invarianz genauer betrachtet. Zuletzt wurde das oben beschriebene und in *Abbildung 33* dargestellte Gesamtmodell mit den Restriktionen zur starken faktoriellen Invarianz und der Modellierung der latenten Differenzvariablen geschätzt. Zur Hypothesenprüfung wurden die Effektstärken der standardisierten Parameterschätzer für die Kovarianz zwischen den latenten Differenzvariablen verglichen.

Die Strukturgleichungsmodelle wurden mithilfe des lavaan-Pakets Version 0.6-12 in R Version 4.0.3 (R Core Team, 2020) anhand der Rohdaten geschätzt. Es wurde der ML-Schätzer gewählt. In die Analysen gingen die Daten von $n = 704$ Fällen ein, die sowohl Teil der Analysestichprobe 1 als auch der Analysestichprobe 2 waren, für die also zu beiden Erhebungszeitpunkten mindestens ein Deutschkompetenzwert vorlag und eine Antwort auf das Standard-Item zum Verstehen der deutschen Sprache, und für die weiterhin zu beiden Messzeitpunkten Antworten zu den Can-Do-Statements vorlagen. Um Fälle mit fehlenden Werten in die Analysen miteinzubeziehen, wurde das FIML-Schätzverfahren angewendet. Für die Modellierung in lavaan habe ich die von Kievit et al. (2018) online zur Verfügung gestellten R-Skripte adaptiert. Die Ergebnisse werden im Folgenden in der Reihenfolge des beschriebenen mehrstufigen Vorgehens berichtet.[10]

Tabelle 19 zeigt, dass alle drei Messmodelle mit unterschiedlichen Restriktionen zur Prüfung der Messinvarianz über die Zeit sehr gut zu den Daten passen. In *Tabelle 20* sind die Ergebnisse der Modellvergleiche dargestellt. Der χ^2-Differenzentest zum Vergleich des Modells schwacher faktorieller Invarianz mit dem Modell konfiguraler Invarianz fällt signifikant aus. Demnach passt das Modell konfiguraler Invarianz besser zu den Daten als das Modell mit den Restriktionen zur schwachen faktoriellen Invarianz, bei dem die Faktorladungen der Indikatoren jeweils über die Zeit gleichgesetzt wurden. Auch beim Vergleich der AIC-Werte schneidet das Modell konfiguraler Invarianz besser ab. Beim Vergleich der BIC-Werte schneiden beide Modelle gleich gut ab, weshalb nach diesem Modellgütekriterium tendenziell das sparsamere Modell schwacher faktorieller Invarianz zu bevorzugen wäre. Insgesamt sind die Ergebnisse des ersten Modellvergleichs somit nicht eindeutig. Da der Fit des Modells schwacher faktorieller Invarianz jedoch sehr gut ist und die

[9] Zur Gleichsetzung der Intercepts der Indikatoren über die Zeit wurde jeweils das Intercept des ersten Indikators auf 0 fixiert und das Intercept des zweiten Indikators gleichgesetzt. Dabei wurden die Intercepts der latenten State-Faktoren zu beiden Messzeitpunkten frei geschätzt (Geiser (2011). Im Gesamtmodell waren die Intercepts der latenten State-Faktoren zum zweiten Messzeitpunkt hingegen auf 0 fixiert, das Intercept der latenten Differenzvariable wurde frei geschätzt.

[10] Die R-Skripte, die die Modellspezifikationen in lavaan enthalten, werden auf Anfrage zur Verfügung gestellt.

Unterschiede in der Modellgüte zwischen beiden Modellen insgesamt relativ gering sind, ist es vertretbar, die Restriktion gleicher Faktorladungen der Indikatoren über die Zeit für weitere Analysen anzuwenden. Der χ^2-Differenzentest zum Vergleich des Modells starker faktorieller Invarianz mit dem Modell schwacher Invarianz fällt ebenfalls signifikant aus. Nach dem χ^2-Differenzentest passt das Modell schwacher faktorieller Invarianz besser zu den Daten als das Modell starker faktorieller Invarianz, bei dem zusätzlich zu den Faktorladungen auch die Intercepts der Indikatoren über die Zeit gleichgesetzt wurden. Beim Vergleich der AIC-Werte schneidet das Modell schwacher faktorieller Invarianz ebenfalls etwas besser ab als das Modell starker faktorieller Invarianz. Nach den BIC-Werten schneidet jedoch das Modell starker faktorieller Invarianz besser ab als das Modell schwacher faktorieller Invarianz. Da auch das Modell starker faktorieller Invarianz eine als sehr gut zu bewertende Modellgüte aufweist und zumindest der BIC-Wert das Modell starker faktorieller Invarianz bevorzugt, ist es vertretbar, die Restriktion gleicher Intercepts der Indikatoren über die Zeit für weitere Analysen anzuwenden. Zusammenfassend ist es nach den Modellvergleichen also nicht eindeutig, ob die Faktoren über die Zeit messinvariant sind. Um die weiteren Analysen durchführen zu können, wird die starke faktorielle Invarianz jedoch angenommen, was zumindest durch die BIC-Werte unterstützt wird.

Tabelle 19. Ausgewählte Indizes der Modellgüte für die Messmodelle mit unterschiedlichen Restriktionen zur Prüfung der Messinvarianz über die Zeit und das Gesamtmodell der Zusammenhänge zwischen latenten Veränderungen verschiedener Kompetenzmaße

Modell	χ^2	*df*	*p*	CFI	RMSEA (90% KI)	SRMR
Modelle zur Prüfung der Messinvarianz						
Modell konfiguraler Invarianz	76.610	22	<.001	.981	.059 [.045, .074]	.020
Modell schwacher faktorieller Invarianz	90.113	24	<.001	.977	.063 [.049, .077]	.028
Modell starker faktorieller Invarianz	98.317	26	<.001	.974	.063 [.050, .076]	.029
Gesamtmodell						
Gesamtmodell	98.317	26	<.001	.974	.063 [.050, .076]	.029

Tabelle 20. Vergleiche der Messmodelle mit unterschiedlichen Restriktionen zur Prüfung der Messinvarianz über die Zeit

Modell	*df*	AIC	BIC	X^2	ΔX^2	Δdf	Δp
Modell konfiguraler Invarianz	22	15 097.52	15 293.46	76.610			
Modell schwacher faktorieller Invarianz	24	15 107.02	15 293.85	90.113	13.504	2	.001

Modell	*df*	AIC	BIC	X^2	ΔX^2	Δdf	Δp
Modell starker faktorieller Invarianz	26	15 111.23	15 288.94	98.317	8.203	2	.017

Unter den Residuen der verschiedenen Indikatoren im Modell starker faktorieller Invarianz traten keine Korrelationen in problematischer Höhe mit Werten über |.10| auf (s. *Tabelle D 5* in Anhang D). In *Tabelle 21* sind die Schätzer der unstandardisierten Faktorladungen und Fehlervarianzen, deren Standardfehler sowie die standardisierten Faktorladungen und Fehlervarianzen der Indikatoren im Messmodell mit Restriktionen zur starken faktoriellen Invarianz gelistet. Die Faktorladungen sind alle hoch. Die Aufklärung der Varianz der verschiedenen Indikatoren durch die Faktoren liegt im Bereich von $R^2 = .56$ bis $R^2 = .86$.

Tabelle 21. ML-Schätzer der Faktorladungen und Fehlervarianzen im Messmodell mit Restriktionen zur starken faktoriellen Invarianz

	Faktorladungen			**Fehlervarianzen**		
	Unstandardisiert		**Standardisiert**	**Unstandardisiert**		**Standardisiert**
Indikator	**Schätzer**	**SE**	**Schätzer**	**Schätzer**	**SE**	**Schätzer**
Tests T1						
Rezeptiver Wortschatz: PPVT-4 T1	1.00	-	.84	0.30	0.03	.29
Grammatikverständnis: TROG-D T1	0.96	0.03	.84	0.28	0.03	.29
Tests T2						
Rezeptiver Wortschatz: PPVT-4 T2	1.00	-	.86	0.26	0.03	.26
Grammatikverständnis: TROG-D T2	0.96	0.03	.85	0.27	0.03	.28
Standard-SE T1						
Standard-Item Verstehen T1	1.00	-	.93	0.06	0.02	.14
Standard-Item Sprechen T1	0.98	0.05	.87	0.11	0.02	.25
Standard-SE T2						
Standard-Item Verstehen T2	1.00	-	.79	0.12	0.01	.37
Standard-Item Sprechen T2	0.98	0.05	.75	0.15	0.01	.44
Can-Do-Statements T1						
Can-Do-Statements T1	1.00	-	.87	0.66	-	.24
Can-Do-Statements T2						
Can-Do-Statements T2	1.00	-	.90	0.52	-	.19

Anmerkungen. T1 = erster Messzeitpunkt; T2 = zweiter Messzeitpunkt; Standard-SE = Standard-Selbsteinschätzung.

Die Schätzer der Faktorvarianzen und -kovarianzen für das Messmodell sind in *Tabelle 22* gelistet. Der Schätzer der Korrelation zwischen den Faktoren der Kompetenztests zu den beiden Messzeitpunkten ist mit .83 sehr hoch, während die Schätzer der Korrelationen zwischen den Faktoren der Standard-Selbsteinschätzungen zu beiden Messzeitpunkten (.37) und zwischen den Can-Do-Statements zu beiden Messzeitpunkten nur moderat ausfallen (.22). Dies spricht dafür, dass die Rangfolge der Teilnehmenden bei den Kompetenztests über die Zeit deutlich stabiler war als bei den beiden Arten von Selbsteinschätzungen. Die Korrelationen zwischen den unterschiedlichen Maßen der Sprachkompetenz zu gleichen Messzeitpunkten liegen in einem moderaten bis hohen Bereich von .32 bis .46. Dies spricht dafür, dass die unterschiedlichen Kompetenzmaße nicht genau dasselbe gemessen haben, sondern neben der Deutschkompetenz von unterschiedlichen Faktoren beeinflusst wurden, wie die Analysen der vorherigen Kapitel bereits gezeigt haben.

Tabelle 22. ML-Schätzer der Faktorvarianzen und Faktorkovarianzen im Messmodell mit Restriktionen zur starken faktoriellen Invarianz

	Unstandardisiert		Standardisiert
Faktor(en)	**Schätzer**	**SE**	**Schätzer**
	Faktorvarianzen		
Tests T1	0.73	0.06	1.00
Tests T2	0.76	0.06	1.00
Standard-SE T1	0.34	0.02	1.00
Standard-SE T2	0.20	0.02	1.00
Can-Do-Statements T1	2.14	0.15	1.00
Can-Do Statements T2	2.21	0.15	1.00
	Faktorkovarianzen		
Tests T1 ↔ Tests T2	0.62	0.05	.83
Tests T1 ↔ Standard-SE T1	0.19	0.02	.38
Tests T1 ↔ Standard-SE T2	0.18	0.02	.46
Tests T1 ↔ Can-Do-Statements T1	0.45	0.06	.36
Tests T1 ↔ Can-Do-Statements T2	0.47	0.06	.37
Tests T2 ↔ Standard-SE T1	0.15	0.02	.29
Tests T2 ↔ Standard-SE T2	0.18	0.02	.46
Tests T2 ↔ Can-Do-Statements T1	0.46	0.06	.36
Tests T2 ↔ Can-Do-Statements T2	0.60	0.06	.46
Standard-SE T1 ↔ Standard-SE T2	0.10	0.01	.37
Standard-SE T1 ↔ Can-Do-Statements T1	0.27	0.04	.32

	Unstandardisiert		Standardisiert
Faktor(en)	**Schätzer**	**SE**	**Schätzer**
Standard-SE T1 ↔ Can-Do-Statements T2	0.12	0.04	.14
Standard-SE T2 ↔ Can-Do-Statements T1	0.19	0.03	.29
Standard-SE T2 ↔ Can-Do-Statements T2	0.27	0.03	.40
Can-Do-Statements T1 ↔ Can-Do-Statements T2	0.47	0.11	.22

Anmerkungen. T1 = erster Messzeitpunkt; T2 = zweiter Messzeitpunkt; Standard-SE = Standard-Selbsteinschätzung.

Als nächstes beschreibe ich die Ergebnisse der Schätzung des Gesamtmodells mit den Restriktionen zur starken faktoriellen Invarianz über die Zeit und der Modellierung der latenten Differenzvariablen. Da es sich bei dem Gesamtmodell mit der Modellierung der latenten Differenzvariablen nur um Reparametrisierungen des Messmodells mit den Restriktionen zur starken faktoriellen Invarianz handelt, ist der Fit der Modelle identisch (Geiser, 2011). Somit sprechen die in *Tabelle 19* berichteten Modellgüteindizes auch hier für eine sehr gute Passung zwischen Modell und Daten. Auch die in *Tabelle D 6* in Anhang D enthaltenen Korrelationen zwischen den Residuen sind sehr gering und damit als unproblematisch zu werten.

Die Faktorladungen der Indikatoren der verschiedenen Kompetenzmaße unterscheiden sich ebenfalls nicht zwischen dem Messmodell und dem Gesamtmodell und sind daher bereits in *Tabelle 21* enthalten. Auch die Varianzaufklärung der manifesten Indikatoren durch die latenten State-Faktoren lag identisch zum Messmodell im Bereich von $R^2 = .56$ bis $R^2 = .86$. Im Gesamtmodell wurden zusätzlich die latenten State-Faktoren zum zweiten Messzeitpunkt jeweils als Indikator der entsprechenden latenten Differenzvariablen modelliert. Die unstandardisierten Parameterschätzer waren jeweils auf 1 fixiert, die standardisierten Schätzer betrugen $\hat{\lambda}_{Tests}^{standardisiert} = 0.57$, $\hat{\lambda}_{Standard-SE}^{standardisiert} = 1.32$ und $\hat{\lambda}_{Can-Do-Statements}^{standardisiert} = 1.24$. Die Varianzaufklärung der latenten State-Faktoren zum zweiten Messzeitpunkt lag entsprechend der Modellspezifikation jeweils bei $R^2 = 1.00$.

Schätzer für die Parameter des strukturellen Teils des Gesamtmodells sind in *Tabelle 23* enthalten. Für die Hypothesenprüfung ist der Vergleich der standardisierten Schätzer der Zusammenhänge zwischen den latenten Differenzvariablen der beiden subjektiven Kompetenzmaße mit der latenten Differenzvariablen der Kompetenztests relevant. Mit einem geschätzten standardisierten Wert von .29 war der Zusammenhang zwischen der latenten Differenzvariable der Can-Do-Statements mit der latenten Differenzvariable der Kompetenztests höher als der Zusammenhang der latenten Differenzvariable der Standard-Selbsteinschätzung mit der latenten Differenzvariable der Kompetenztests, welcher auf einen standardisierten Wert von .17 geschätzt wurde. Die latent modellierten Differenzen der Can-Do-Statements hingen demnach stärker mit den latent modellierten Differenzen der Kompetenztests zusammen als die latent modellierten Differenzen der Standard-

Selbsteinschätzungen. Dieser Befund unterstützt Hypothese 5b, wonach die Can-Do-Statements Veränderungen der Kompetenzen besser abbilden als die Standard-Selbsteinschätzungen, gemessen an den Kompetenztests als Kriterium.

Tabelle 23. ML-Schätzer für die Faktorvarianzen, direkte Effekte zwischen Faktoren und Faktorkovarianzen im Gesamtmodell der Zusammenhänge zwischen latenten Veränderungen verschiedener Kompetenzmaße

	Unstandardisiert		Standardisiert
Faktor(en)	**Schätzer**	**SE**	**Schätzer**
Faktorvarianzen			
Tests T2	0.00	-	.00
Differenz Tests	0.23	0.03	.91
Tests T1	0.73	0.06	1.00
Standard-SE T2	0.00	-	.00
Differenz Standard-SE	0.15	0.01	.42
Standard-SE T1	0.34	0.02	1.00
Can-Do-Statements T2	0.00	-	.00
Differenz Can-Do-Statements	1.89	0.13	.55
Can-Do-Statements T1	2.14	0.15	1.00
Direkte Effekte der Faktoren zu Messzeitpunkt 1 auf Faktoren zu Messzeitpunkt 2 oder Differenzvariablen			
Tests T1 → Tests T2	1.00	-	0.98
Tests T1 → Differenz Tests	–0.16	0.05	–.27
Standard-SE T1 → Standard-SE T2	1.00	-	1.31
Standard-SE T1 → Differenz Standard-SE	–0.84	0.04	–.84
Can-Do-Statements T1 → Can-Do-Statements T2	1.00	-	0.98
Can-Do-Statements T1 → Differenz Can-Do-Statements	–0.90	0.05	–.71
Standard-SE T1 → Differenz Tests	–0.08	0.06	–.10
Can-Do-Statements T1 → Differenz Tests	0.05	0.02	.15
Tests T1 → Differenz Standard-SE	0.18	0.03	.26
Can-Do-Statements T1 → Differenz Standard-SE	0.03	0.02	.08
Tests T1 → Differenz Can-Do-Statements	0.60	0.09	.28
Standard-SE T1 → Differenz Can-Do-Statements	–0.05	0.12	–.02
Kovarianzen zwischen Faktoren zu Messzeitpunkt 1 oder zwischen Differenzvariablen			
Tests T1 ↔ Standard-SE T1	0.19	0.02	.38
Tests T1 ↔ Can-Do-Statements T1	0.45	0.06	.36

	Unstandardisiert		Standardisiert
Faktor(en)	**Schätzer**	**SE**	**Schätzer**
Standard-SE T1 ↔ Can-Do-Statements T1	0.27	0.04	.32
Differenz Tests ↔ Differenz Standard-SE	0.03	0.01	.17
Differenz Tests ↔ Differenz Can-Do-Statements	0.19	0.04	.29
Differenz Standard-SE ↔ Differenz Can-Do-Statements	0.15	0.03	.28

Anmerkungen. T1 = erster Messzeitpunkt; T2 = zweiter Messzeitpunkt; Standard-SE = Standard-Selbsteinschätzung.

5 Diskussion

Gegenstand dieser Arbeit sind die Selbsteinschätzungen von sprachlichen Kompetenzen im Deutschen jugendlicher Flüchtlinge. Es wurde untersucht, wie genau die jugendlichen Flüchtlinge ihre Deutschkompetenzen einschätzen, wobei nicht nur die Korrelation zwischen Selbsteinschätzungen und objektiven Kompetenzmaßen analysiert wurde, sondern auch das Ausmaß der allgemeinen Verzerrung und inwiefern die Selbsteinschätzungen die Variation der tatsächlichen Kompetenz abbilden können. Zudem wurden Faktoren identifiziert, die die Selbsteinschätzungen systematisch beeinflussen, unabhängig von der tatsächlichen Deutschkompetenz, die mit den Selbsteinschätzungen erfasst werden soll. Zuletzt wurden verschiedene Arten von Selbsteinschätzungsitems miteinander verglichen und geprüft, ob allgemein formulierte Selbsteinschätzungsitems oder spezifischere, auf konkrete Situationen bezogene Items die Veränderung von Deutschkompetenzen zwischen zwei Messzeitpunkten besser erfassen können.

Im Folgenden werden zunächst die Ergebnisse der Genauigkeit der Selbsteinschätzungen (s. Kapitel 5.1), der Faktoren, die die Selbsteinschätzungen beeinflussen (s. Kapitel 5.2) und des Vergleichs der verschiedenen Selbsteinschätzungsitems (s. Kapitel 5.3) jeweils einzeln zusammengefasst und diskutiert und es wird auf spezifische Implikationen und Limitationen eingegangen. Anschließend werden in Kapitel 5.4 allgemeine Limitationen, die sich auf mehrere Analysen beziehen, diskutiert. Daraufhin wird in Kapitel 5.5 darauf eingegangen, welche Konsequenzen die Messung von Sprachkompetenzen mittels Selbsteinschätzungen mit sich bringen und wie die Genauigkeit von Selbsteinschätzungen optimiert werden könnte. Die Diskussion endet mit einem Fazit (s. Kapitel 5.6).

5.1 Genauigkeit der Selbsteinschätzungen

Da Selbsteinschätzungen eine ökonomischere Möglichkeit darstellen, deutsche Sprachkompetenzen jugendlicher Flüchtlinge zu erfassen, als Deutschkompetenztests, war es ein Ziel dieser Arbeit, die Genauigkeit der Selbsteinschätzungen und deren Eignung als Sprachkompetenzmaß zu analysieren. Dazu wurde zum einen analysiert, wie gut die Selbsteinschätzungen zwischen den Teilnehmenden hinsichtlich ihrer Deutschkompetenz diskriminieren. Es hat sich gezeigt, dass die Diskrimination der Kompetenzen anhand der Selbsteinschätzungen besser ist, als es in Hypothese 1a vorhergesagt wurde. Die Zusammenhänge zwischen den Selbsteinschätzungen und objektiven Kompetenzmaßen wiesen eine mittlere Effektstärke auf, vorhergesagt wurde eine kleine Effektstärke. Die Korrelationen mittlerer Effektstärke zwischen Selbsteinschätzungen und Kompetenztests bedeuten dennoch, dass nur ein sehr geringer Anteil der Varianz der

Selbsteinschätzungen durch die tatsächlichen Kompetenzen erklärt werden kann und die Selbsteinschätzungen kein konvergent valides Maß der Sprachkompetenz darstellen und diese nur relativ ungenau erfassen können.

Da nur ein kleiner Teil der Varianz durch die tatsächliche Kompetenz erklärt werden kann, wird ein großer Anteil der Varianz der Selbsteinschätzungen durch verschiedene mehr oder weniger systematische Fehlerquellen beeinflusst. Mögliche Ursachen für Ungenauigkeiten bei der Selbsteinschätzung wurden in Kapitel 1 zum theoretischen und empirischen Hintergrund herausgearbeitet: Ungenauigkeiten können sowohl im Prozess der Urteilsbildung als auch bei der Beantwortung der Selbsteinschätzungsfragen entstehen. In der Terminologie des RAM (Funder, 1995) beschrieben ist, wie in Abschnitt 2.1.1 argumentiert, davon auszugehen, dass den jugendlichen Flüchtlingen genügend relevante Informationen zu ihrer Deutschkompetenz zur Verfügung stehen. Es ist wahrscheinlich, dass es im Prozess der Urteilsbildung zur deutschen Sprachkompetenz insbesondere auf der letzten Stufe (Utilization) zu Ungenauigkeiten kommt, weil die Informationen teilweise nicht richtig interpretiert und verarbeitet werden. Dazu gehört, dass unterschiedliche internale und externale Referenzrahmen die Selbsteinschätzungen unterschiedlich beeinflussen (s. Abschnitt 1.2.4 und Abschnitt 1.2.5), und dass Personen unterschiedlich stark zum Self-Enhancement neigen (s. Kapitel 1.3).

Eine weitere Ursache der Ungenauigkeit stellt der nicht im RAM enthaltene Schritt der Auswahl einer passenden Antwortkategorie dar. Für diesen letzten Schritt ist insbesondere relevant, ob die Teilnehmenden sowohl die Frage als auch die Antwortskala einheitlich interpretieren und darin mit den Forschenden übereinstimmen (Krosnick & Presser, 2010; vgl. Abschnitt 1.7.2). Trotz der Vorgabe der Dimensionen Verstehen und Sprechen ist der Kompetenzbereich bei den Standard-, den Schieberegler- und den Vergleich-Items nicht eindeutig definiert und Teilnehmende können ihre Einschätzung auf unterschiedliche Bereiche fokussiert haben, wie z.B. Wortschatz oder Grammatik, was die Vergleichbarkeit der Einschätzungen beeinträchtigt (z.B. Dunning et al., 1989; vgl. Abschnitt 1.3.1).

Darüber hinaus ist insbesondere für subjektive Skalen wie die Standard- oder Schieberegler-Items nicht gegeben, dass die Teilnehmenden die Antwortskalen einheitlich interpretieren, denn die Items geben keine Referenz dafür vor, was es z.B. bedeutet, sehr gut Deutsch zu sprechen. Nach dem Shifting-Standards-Modell gehen Urteilende davon aus, dass die Punkte der Ratingskala die Verteilung der Gruppe wiedergeben, der die zu beurteilende Person angehört (z.B. Biernat, 2005; vgl. Abschnitt 1.2.4). Nach dem BFLPE beeinflusst das Kompetenzniveau in der direkten Umwelt einer Person die Selbsteinschätzungen von Kompetenzen (Marsh, 1987; vgl. Abschnitt 1.2.4; Marsh & Parker, 1984). Das direkte Umfeld, aus dem die Teilnehmenden ihre Vergleichsinformationen beziehen können, ist jedoch immer nur eine sehr kleine selektive Stichprobe, die sich zwischen Teilnehmenden unterscheidet, da diese individuell für die Teilnahme an der Studie

ausgewählt wurden und in der Regel keine gemeinsame direkte Umwelt miteinander teilen. Das bedeutet, dass die Teilnehmenden die Punkte der subjektiven Ratingskalen unterschiedlich interpretieren und die Interpretation nur selten dem entsprechen kann, was in dieser Arbeit als genaue Selbsteinschätzung definiert wurde. Die Definition genauer Selbsteinschätzungen wurde in dieser Arbeit aus den Informationen zu den Deutschkompetenzen großer repräsentativer Stichproben abgeleitet (vgl. Abschnitt 3.3.2.3), welche den Teilnehmenden für ihre Skaleninterpretation hingegen nicht zur Verfügung standen. Für die Vergleich-Items wurde zwar eine Referenzgruppe vorgegeben, diese scheint jedoch keine angemessene Vergleichsgruppe zu sein, wie in Kapitel 5.3 diskutiert wird. Die Can-Do-Statements beziehen sich auf spezifischere Situationen und stellen objektivere Items dar, werden aber wahrscheinlich trotzdem unterschiedlich interpretiert. Die uneinheitliche Interpretation der Selbsteinschätzungsitems und deren Antwortskalen stellt also eine Ursache für die Ungenauigkeit der Selbsteinschätzungen dar.

Weiterhin kann auch das junge Alter der Teilnehmenden dazu beigetragen haben, dass keine hohe Validität der Selbsteinschätzungen erreicht wurde. Bei Personen im mittleren Jugendalter ist das Selbstbild insgesamt noch nicht gefestigt (vgl. Harter, 2012; vgl. Kapitel 1.4) und in der Folge müssen Selbsteinschätzungen häufiger ad hoc gebildet werden. Dieser Prozess wird von der aktuellen Zugänglichkeit von Informationen beeinflusst und ist damit ungenau (vgl. Strack & Martin, 1987; vgl. Abschnitt 1.7.1). Diese Annahme wird dadurch untermauert, dass die Selbsteinschätzungen in der siebten Erhebungswelle, in der die Jugendlichen durchschnittlich das höhere Jugendalter erreicht haben, tendenziell stärker mit den objektiven Kompetenzmaßen korrelierten als in der ersten Erhebungswelle. Außerdem waren die Korrelationen der Selbsteinschätzungen mit den objektiven Maßen in der ersten Erhebungswelle ähnlich hoch wie die der sich ebenfalls im mittleren Jugendalter befindenden Neuntklässlerinnen und Neuntklässler mit Migrationshintergrund im NEPS, welche auch mit den Standard-Items befragt wurden (Edele et al., 2015). Das junge Alter der Teilnehmenden und das damit einhergehende noch nicht gefestigte Selbstbild können dazu beigetragen haben, dass die Selbsteinschätzungen kurzfristig vorgenommen wurden und deshalb ungenauer waren, als es bei Stichproben mit erwachsenen Teilnehmenden oft der Fall ist (z.B. Edele et al., 2015; vgl. Kapitel 1.9).

Für die Wahl der Kriteriumsmaße wird angenommen, dass diese die Korrelationen günstig beeinflusst haben. Die Passung zwischen selbsteingeschätzter und getesteter Kompetenz ist in dieser Arbeit für die Verstehen-Items sehr gut und sowohl das Grammatikverständnis als auch der Wortschatz wurden mit etablierten Kompetenztests hoher Qualität gemessen. Obwohl auch die Ergebnisse von Kompetenztests immer Messfehlern unterliegen, ist davon auszugehen, dass Messfehler die Korrelationen nicht stärker beeinträchtigt haben, als es in anderen Studien der Fall ist, in denen die Korrelationen teilweise sehr hoch sind (vgl. Kapitel 1.9).

Zusammenfassend waren die Korrelationen der Selbsteinschätzungen mit den Kompetenztests höher als erwartet, aber im Sinne der konvergenten Validität dennoch als niedrig zu bewerten. Die niedrige Validität wird damit erklärt, dass die im RAM beschriebene Urteilsbildung mit Verzerrungen einhergeht und dass uneinheitliche Interpretationen von Selbsteinschätzungsfragen und Antwortskalen die Vergleichbarkeit der Selbsteinschätzungen beeinträchtigen. Als Besonderheit der untersuchten Gruppe hat sich wahrscheinlich das junge Alter negativ auf die Genauigkeit der Selbsteinschätzungen ausgewirkt.

Um die Genauigkeit der Selbsteinschätzungen zu beurteilen, habe ich neben den Korrelationen zwischen Selbsteinschätzungen und Deutschkompetenztests auch untersucht, inwiefern die Teilnehmenden ihre Deutschkompetenzen im Durchschnitt über- oder unterschätzen. Insgesamt gibt es deutliche Hinweise, dass die jugendlichen Flüchtlinge ihre sprachlichen Kompetenzen im Deutschen im Durchschnitt überschätzen, womit Hypothese 1b bestätigt wurde.

Dass die Teilnehmenden ihre sprachliche Kompetenz im Deutschen durchschnittlich überschätzen, ist nach den Ausführungen in Kapitel 1 und Abschnitt 2.1.2 auf die allgemeine Neigung zum Self-Enhancement zurückzuführen, welche durch verschiedene Faktoren begünstigt wird. Dazu gehört, dass deutsche Sprachkompetenzen wichtig für die in Deutschland lebenden jugendlichen Flüchtlinge sind und entsprechend dem Self-Centrality-Breeds-Self-Enhancement-Prinzip die Einschätzung von persönlich wichtigen Kompetenzen besonders durch das Self-Enhancement-Motiv beeinflusst wird (Gebauer et al., 2013; Sedikides & Alicke, 2019; vgl. Abschnitt 1.3.1). Weiterhin wird das Self-Enhancement begünstigt, weil die Kompetenz verschiedene Kompetenzbereiche umfasst und Interpretationsspielraum lässt (Dunning et al., 1989) und weil die Teilnehmenden keine Konsequenzen im Zusammenhang mit der Genauigkeit ihrer Einschätzung zu erwarten hatten (Sedikides et al., 2002; vgl. Abschnitt 1.3.3). Zudem waren die Teilnehmenden durch die lange Befragungssituation kognitiver Belastung ausgesetzt, was eine heuristische Beantwortung der Selbsteinschätzungen mit Neigung zum Self-Enhancement begünstigt (z.B. Alicke et al., 1995; vgl. Abschnitt 1.3.1.4 und Abschnitt 1.3.2).

Hinsichtlich des kulturellen Einflusses auf die Selbsteinschätzungen hat sich bestätigt, dass die Teilnehmenden nicht zu einem rein interdependenten Selbstbild mit Hang zur Bescheidenheit neigen, sondern dass die jugendlichen Geflüchteten trotz ihrer Herkunft aus Ländern mit eher kollektivistischer Kultur zum Self-Enhancement neigen. Das könnte bedeuten, dass entweder die Neigung zum Self-Enhancement entgegen der Ausführungen in Kapitel 1.5 auch in den Herkunftsländern der Teilnehmenden insgesamt verbreitet ist, oder dass die Jugendlichen durch ihren Aufenthalt in Deutschland eine individualistische Prägung erfahren haben und deshalb so stark zum Self-Enhancement neigen.

Neben den genannten vorrangig motivationalen Faktoren, die die Selbstüberschätzung begünstigt haben, ist auch davon auszugehen, dass Teilnehmende, wie im Dunning-Kruger-Effekt beschrieben, eigene Fehler aufgrund mangelnder Fähigkeiten nicht immer erkennen können und ihre Kompetenzen deshalb überschätzen (Kruger & Dunning, 1999; vgl. Abschnitt 1.6.2.1).

Ein weiterer Faktor, der bereits als Grund für die Ungenauigkeit der Selbsteinschätzungen diskutiert wurde und der auch die gefundene durchschnittliche Überschätzung begünstigt haben könnte, ist dass die Teilnehmenden die Skalenpunkte möglicherweise anders interpretiert haben, als die in dieser Arbeit getroffenen Annahmen für die Definition einer genauen Selbsteinschätzung intendieren. Für die Standard-Items und die Schieberegler-Items wurde angenommen, dass die Skala die Verteilung der jugendlichen Flüchtlinge der hier betrachteten Stichprobe darstellt und eine im Vergleich zu dieser Gruppe durchschnittliche Leistung bei den Standard-Items der Mitte zwischen den Kategorien *eher gut* und *eher schlecht* entspricht, und bei den Schieberegler-Items der Mitte der Skala (vgl. Abschnitt 3.3.2.3 und Abschnitt 4.4.2). Womöglich vergleichen sich die jugendlichen Flüchtlinge jedoch nicht nur mit anderen jugendlichen Flüchtlingen, sondern z.B. auch mit ihren Eltern. Da die Jugendlichen über ihren Schulbesuch regelmäßig Kontakt zur deutschen Sprache haben, was auf die Eltern wahrscheinlich nicht immer in demselben Ausmaß zutrifft, ist anzunehmen, dass dieser Vergleich häufig positiv ausfällt und die positiven Einschätzungen somit zusätzlich erklären könnte. Tatsächlich haben geflüchtete Personen im Erwachsenenalter in der IAB-BAMF-SOEP-Befragung von Geflüchteten, die zwischen 2013 und 2016 nach Deutschland gekommen sind und zum Zeitpunkt der Befragung im Jahr 2019 mehrheitlich seit vier Jahren in Deutschland lebten, ihre Deutschkompetenzen niedriger selbsteingeschätzt als die Jugendlichen der ReGES-Stichprobe (Niehues et al., 2021). In der IAB-BAMF-SOEP-Befragung von Geflüchteten gaben 13% der Stichprobe sehr gute Deutschkenntnisse an, 34% gaben gute Deutschkenntnisse an, 35% wählten die Kategorie *es geht* und 18% wählten die Kategorien *eher schlecht* oder *gar nicht*. Auch die Erwachsenen haben jedoch überwiegend positive Kategorien gewählt, was unter der Annahme, dass die Skala der Verteilung der jeweiligen Gesamtstichprobe entsprechen sollte, ebenfalls auf eine Überschätzung hindeutet, jedoch war dies bei den Erwachsenen in einem geringeren Ausmaß der Fall als bei den Jugendlichen. Geht man also davon aus, dass die Jugendlichen tatsächlich über bessere Deutschkompetenzen verfügen als die Erwachsenen, würde dies dafürsprechen, dass die hohen Selbstüberschätzungen der Jugendlichen neben der bei beiden Gruppen gefundenen Tendenz zum Self-Enhancement auch darauf zurückzuführen sind, dass die Teilnehmenden nicht nur die eigene Stichprobe als Vergleichsgruppe heranziehen, sondern auch Personen aus anderen Altersgruppen.

Zusammenfassend haben die jugendlichen Flüchtlinge ihre Deutschkompetenzen erwartungsgemäß im Durchschnitt überschätzt, was auf die Neigung zum Self-Enhancement, auf den Dunning-Kruger-Effekt sowie auf die Interpretation der Antwortskala durch die Teilnehmenden

zurückzuführen ist. Eine dem Self-Enhancement entgegengesetzte Bescheidenheit bei der Beantwortung der Selbsteinschätzungen, wie sie in kollektivistisch geprägten Kulturen häufig gefunden wurde, haben die jugendlichen Flüchtlinge nicht gezeigt.

Als letzte Komponente der Genauigkeit der Selbsteinschätzungen wurde die Variation der Selbsteinschätzungen mit der Variation der tatsächlichen Kompetenz verglichen. Es hat sich gezeigt, dass die Selbsteinschätzungen die Variation der Kompetenzen nicht angemessen abbilden, was Hypothese 1c bestätigt. Die Variation der Selbsteinschätzungen mit fünf- bis elfstufiger Antwortskala ist geringer als die Variation der objektiven Kompetenzmaße. Zunächst ist dies darauf zurückzuführen, dass die Anzahl der Kategorien der Selbsteinschätzungen deutlich geringer ist als die Anzahl der Items der Kompetenztests, die zu einem Summenwert verrechnet werden. Da Sprachkompetenzen ein sehr breites Kontinuum umfassen, von Personen ohne jegliche Kenntnisse bis hin zu einem sehr guten muttersprachlichen Niveau, scheint eine Skala mit nur fünf Kategorien sehr knapp zu sein, um die Variation angemessen zu erfassen.

Tatsächlich hat sich zumindest bei den Standard-Items ein Deckeneffekt gezeigt, was bedeutet, dass den Teilnehmenden eine noch bessere Kategorie gefehlt hat, um zwischen den Personen im oberen Leistungsbereich zu differenzieren. Darüber hinaus wird die Variation der Selbsteinschätzungen eingeschränkt, weil die Teilnehmenden die wenigen positiven Kategorien bevorzugt wählen und sich die Selbsteinschätzungen somit auf noch weniger Kategorien verteilen. Wurden mehr Kategorien angeboten, wie bei den Schieberegler-Items, verteilten sich die Antworten auch auf mehr Kategorien, konzentrierten sich aber auch hier auf die obere Hälfte der Skala. Demnach gibt es kaum Teilnehmende, deren Kompetenz laut Selbsteinschätzung im sehr schlechten Bereich liegt, was je nach Definition einer sehr schlechten Kompetenz nicht der tatsächlichen Kompetenzverteilung entspricht. Es ist weiterhin möglich, dass die angebotenen Antwortkategorien den mittleren Kompetenzbereich nicht angemessen wiedergeben und die Teilnehmenden z.B. bei den Standard-Items eine mittlere Kategorie vermissen und sich schwertun, sich zwischen den Kategorien *eher schlecht* und *eher gut* zu entscheiden. Das heißt auch im mittleren Bereich wird die Variation der Kompetenzen durch die Selbsteinschätzungen nicht angemessen erfasst. Der Summenwert der Can-Do-Statements hat eine etwas größere Variationsbreite als die Standard- und die Vergleich-Items. Aber auch hier kann die tatsächliche Variation der Kompetenzen nicht angemessen abgebildet werden. Da ein Großteil der Items in der Stichprobe von sehr vielen Personen gewählt wurde und somit eine geringe Schwierigkeit aufwies, gab es auch hier einen deutlichen Deckeneffekt.

Zusammenfassend ist die Variation der Selbsteinschätzungen deutlich geringer als die der tatsächlich zugrundeliegenden Kompetenz, was auf die geringe Anzahl an Antwortkategorien zurückzuführen ist und darauf, dass die Teilnehmenden nur einen Teil der verfügbaren Antwortkategorien für ihre Selbsteinschätzungen genutzt haben.

5.2 Einflussfaktoren der Selbsteinschätzungen

Damit Auswirkungen der Verwendung von Selbsteinschätzungen als Deutschkompetenzmaß eingeschätzt werden können, wurden zudem Einflussfaktoren von Selbsteinschätzungen identifiziert. In den hier durchgeführten Analysen wurden die möglichen Einflussfaktoren gemeinsam in ein Strukturgleichungsmodell aufgenommen, mit dem Ziel, die systematische Varianz der Selbsteinschätzungen zu erklären. So konnte unter Kontrolle der jeweils anderen Einflussfaktoren gezeigt werden, wie sich die Einflussfaktoren jeweils auf die Selbsteinschätzungen auswirken. Die Ergebnisse haben gezeigt, dass neben der zugrundeliegenden Kompetenz nicht nur unsystematische Messfehler die Varianz der Selbsteinschätzungen erklären, sondern bestimmte Faktoren die Selbsteinschätzungen systematisch beeinflussen. Zum einen haben Personen mit höheren Fähigkeiten zum schlussfolgernden Denken ihre Deutschkompetenzen erwartungsgemäß niedriger eingeschätzt als Personen mit niedrigeren Fähigkeiten zum schlussfolgernden Denken unter Kontrolle der objektiv gemessenen Kompetenz (s. Hypothese 2c). Weiterhin haben Teilnehmende, die mehr Engagement beim Deutschlernen gezeigt haben, ihre Deutschkompetenzen erwartungsgemäß höher eingeschätzt als Personen, die weniger Engagement beim Deutschlernen gezeigt haben, unter Kontrolle der objektiv gemessenen Kompetenz (s. Hypothese 2b). Für die schulische Leistung in Mathematik (s. Hypothese 2a) und die Teilnahme an einem Deutschkurs (s. Hypothese 2d) und Deutschtest (s. Hypothese 2e) konnte kein Effekt in die erwartete Richtung auf die Selbsteinschätzung der Deutschkompetenz gefunden werden.

Wie in Kapitel 2.2 erläutert, sind die mit höheren Fähigkeiten zum schlussfolgernden Denken einhergehenden niedrigeren Selbsteinschätzungen vermutlich darauf zurückzuführen, dass die komplexen Anforderungen des Urteilsprozesses mithilfe von höheren Fähigkeiten zum schlussfolgernden Denken besser gemeistert werden können und dadurch das Ausmaß der Überschätzung reduziert wird. Das unterstützt die Annahme des RAMs, dass bestimmte Fähigkeiten urteilender Personen, und zwar insbesondere kognitive Fähigkeiten, einen Moderator der Urteilsgenauigkeit darstellen (z.B. Funder, 1999; vgl. Abschnitt 1.6.2.1).

Dass Personen, die mehr Engagement beim Deutschlernen gezeigt haben, ihre Deutschkompetenzen höher einschätzen als Personen, die weniger Engagement beim Deutschlernen gezeigt haben, unterstützt die Argumentation in Kapitel 2.2, dass denjenigen Personen, die mehr Engagement zeigen, die sprachlichen Kompetenzen im Deutschen auch persönlich wichtiger sind und nach dem Self-Centrality-Breeds-Self-Enhancement-Prinzip diese Personen ihre Deutschkompetenzen stärker überschätzen (z.B. Gebauer et al., 2013; vgl. Abschnitt 1.3.1). Die gefundene Effektstärke ist jedoch gering, möglicherweise weil die sprachlichen Kompetenzen im Deutschen den

meisten Teilnehmenden sehr wichtig sind und die individuelle Varianz sich darüber hinaus nur noch geringfügig auf die Selbsteinschätzungen auswirkt.

Dass die Leistung in Mathematik nicht den aufgrund des I/E-Modells erwarteten negativen Effekt (vgl. Kapitel 2.2 und Abschnitt 1.2.5) auf die Selbsteinschätzungen hatte, könnte auf verschiedene Gründe zurückzuführen sein. Zum einen könnten die herangezogenen Selbstberichte der Mathematiknoten ungenau sein, da die Noten in der Befragungssituation möglicherweise falsch erinnert wurden (vgl. Abschnitt 3.3.3). Auch bei den Angaben der Lehrerinnen und Lehrer ist dies möglich, wobei hier eher davon ausgegangen werden kann, dass die Lehrerinnen und Lehrer beim Ausfüllen des Papierfragebogens die Noten der Schülerinnen und Schüler nachgeschaut und dann genau beantwortet haben. Fraglich ist darüber hinaus, wie valide und reliabel die Mathenoten die Leistung in Mathematik wiedergeben. Die jugendlichen Flüchtlinge besuchten zum Zeitpunkt der Benotung erst seit Kurzem eine deutsche Schule und es ist zu erwarten, dass sie vorausgehende Inhalte des Lehrplans verpasst haben, die möglicherweise relevant für den weiteren Wissenserwerb gewesen wären. Außerdem hatten sie bei mangelnden Sprachkompetenzen Nachteile beim Erwerb des Lehrstoffs. Eine angemessene Beurteilung der Leistung in Mathematik war aufgrund dieser Umstände entsprechend schwierig und wurde von Lehrkräften wahrscheinlich unterschiedlich gehandhabt, was die Vergleichbarkeit der Mathematiknoten zwischen den Teilnehmenden, die unterschiedliche Schulen besuchten und von unterschiedlichen Lehrkräften bewertet wurden, beeinträchtigt. Außerdem haben die Jugendlichen selbst die Aussagekraft der Mathematiknote aus diesen Gründen womöglich abgewertet und die Bewertungen ihrer Leistung weniger verinnerlicht und ein entsprechend unstabiles Selbstbild hinsichtlich der Mathematikleistung ausgebildet als Teilnehmende von Studien zum I/E-Modell. Entsprechend haben sie internale Vergleiche mit ihrem mathematischen Selbstbild vermutlich weniger für die Beurteilung ihrer sprachlichen Kompetenzen im Deutschen herangezogen, als aufgrund des I/E-Modells erwartet wurde. Möglicherweise kam es im Gegenteil sogar zu Assimilationseffekten, wenn die jugendlichen Flüchtlinge davon ausgingen, dass die Mathematikleistung insbesondere durch die Deutschkompetenzen beeinflusst wird, da mangelnde Deutschkompetenzen auch das Verständnis des Mathematikunterrichts einschränken. Assimilationseffekte treten dann auf, wenn sich die Leistung in einem Fach positiv auf das Selbstkonzept in einem anderen, ähnlichen Fach auswirkt und wenn Personen davon ausgehen, dass die Fähigkeiten, die für die Leistungen in den beiden Fächern relevant sind, ähnlicher sind, als es tatsächlich der Fall ist (Helm et al., 2020; Möller et al., 2015; vgl. Abschnitt 1.2.5).

Für die Teilnahme an einem Deutschkurs wurde erwartet, dass diese Selbstüberschätzungen reduziert, indem die Teilnehmenden auf Fehler aufmerksam gemacht werden, die sie, wie im Dunning-Kruger-Effekt beschrieben, sonst nicht wahrgenommen hätten (vgl. Kapitel 2.2). Für die Teilnahme an einem Deutschkurs zeigte sich jedoch ein gegenteiliger Effekt: Teilnehmende, die an

einem Deutschkurs teilgenommen haben, haben ihre Deutschkompetenzen stärker überschätzt als Teilnehmende, die an keinem Deutschkurs teilgenommen haben. Das könnte bedeuten, dass das Feedback in den Sprachkursen nicht die vorhergesagte Wirkung auf die Selbsteinschätzungen der Teilnehmenden hatte, oder dass in den Sprachkursen weniger korrigierendes Feedback als vielmehr bestärkendes Feedback gegeben wird, was den positiven Effekt der Teilnahme an einem Sprachkurs auf die Höhe der Selbsteinschätzung erklären könnte. Problematisch an der hier durchgeführten Untersuchung des Einflusses der Teilnahme an einem Deutschkurs auf die Selbsteinschätzungen der sprachlichen Kompetenzen im Deutschen ist, dass nur die Information herangezogen wurde, ob Teilnehmende zum Zeitpunkt der Befragung oder zu einem früheren Zeitpunkt an einem Deutschkurs teilnehmen bzw. teilgenommen haben. Es lagen jedoch keine weiteren Informationen zu den Sprachkursen vor, wie z.B. der Umfang und die Inhalte des Sprachkurses, das Niveau des Sprachkurses, die Ausbildung der Lehrkräfte oder das Ausmaß von und der Umgang mit Feedback im Kurs. Um ein besseres Verständnis für die Auswirkung des Besuchs eines Deutschkurses und insbesondere von Feedback im Rahmen von Deutschkursen auf die Selbsteinschätzungen zu erlangen, müssten systematische Untersuchungen durchgeführt werden, in denen solche Aspekte gezielt manipuliert oder kontrolliert werden.

Weiterhin wurde für die Teilnahme an einem offiziellen Deutschtest ein negativer Effekt auf die Selbsteinschätzungen vorhergesagt, welcher sich nicht bestätigt hat. Offizielle Deutschtests bieten einen Referenzrahmen, anhand dessen die eigene Sprachkompetenz eingestuft werden kann und somit Selbstüberschätzungen reduziert werden könnten (vgl. Kapitel 2.2). Zu welchem Zeitpunkt der Test durchgeführt wurde, wurde jedoch nicht berücksichtigt und möglicherweise waren die Tests in einigen Fällen schon länger her und das Ergebnis veraltet und den Teilnehmenden nicht mehr präsent. Außerdem könnte es sein, dass die Teilnehmenden nicht ausreichend aufgeklärt wurden über die Bedeutung ihres Ergebnisses und die Einstufung in einen größeren Referenzrahmen und dass sie die Information deshalb für die Selbsteinschätzung nicht angemessen nutzen konnten. Um festzustellen, ob die Teilnahme an einem Deutschtest also tatsächlich keine Auswirkung auf die Höhe der Selbsteinschätzung hat, sollten Studien durchgeführt werden, in denen genauere Informationen zum Zeitpunkt der Testung und zur Art der Rückmeldung kontrolliert werden. Beispielsweise könnten sich kurz zurückliegende Tests, bei denen die Teilnehmenden eine ausführliche Rückmeldung zu ihrem Ergebnis und dessen Einordnung erhalten haben, dennoch wie erwartet auf die Selbsteinschätzungen auswirken.

Insgesamt lässt sich festhalten, dass die tatsächliche zugrundeliegende Deutschkompetenz, die Fähigkeit zum schlussfolgernden Denken und das Engagement beim Deutschlernen einen relevanten Anteil der Varianz der Selbsteinschätzungen erklären. Damit bestätigt sich, dass Funders (1995) Good-Judge-Moderator auf Selbsteinschätzungen von Kompetenzen übertragen werden

kann, weil abhängig von bestimmten Eigenschaften Personen ihre Kompetenzen besser einschätzen können als andere.

Drei einschränkende Aspekte bleiben in Bezug auf die hier durchgeführten Analysen zu den Einflussfaktoren von Selbsteinschätzungen zu diskutieren. Zum einen lassen die Ergebnisse keine kausalen Schlüsse zu, sodass insbesondere für den Einfluss des Engagements beim Deutschlernen auf die Selbsteinschätzungen auch ein umgekehrter Effekt nicht auszuschließen ist, der den gefundenen Zusammenhang erklären könnte. Es ist also auch möglich, dass sich die Selbsteinschätzung der sprachlichen Kompetenz im Deutschen positiv darauf auswirkt, wieviel Engagement die jugendlichen Flüchtlinge beim Deutschlernen zeigen.

Zum anderen wurde in dieser Arbeit untersucht, inwiefern sich verschiedene Faktoren auf die Höhe der Selbsteinschätzungen auswirken. Für manche Einflussfaktoren sind jedoch im Fall von Überschätzungen Effekte auf die Selbsteinschätzung in die andere Richtung zu erwarten als im Fall von Unterschätzungen, weil sie sich nicht nur auf die Höhe der Selbsteinschätzung auswirken, sondern auf die Genauigkeit der Selbsteinschätzungen, sodass Unterschätzungen ebenso reduziert oder erhöht werden wie Überschätzungen. Dies betrifft z.B. die Hypothesen zum Einfluss von Fähigkeiten zum schlussfolgernden Denken und zur Teilnahme an einem Deutschtest. Aus theoretischer Sicht ist nicht nur zu erwarten, dass höhere Ausprägungen der Fähigkeit zum schlussfolgernden Denken bzw. die Teilnahme an einem Deutschtest bei Fällen, die ihre Deutschkompetenzen überschätzen einen negativen Effekt auf die Höhe der Selbsteinschätzung haben, sondern auch, dass sie bei Fällen, die ihre Deutschkompetenzen unterschätzen einen positiven Effekt auf die Höhe der Selbsteinschätzung haben. Unter der Annahme, dass die Teilnehmenden ihre Deutschkompetenzen in den meisten Fällen überschätzt haben, habe ich jedoch auch für den Einfluss der betroffenen Variablen auf die Höhe der Selbsteinschätzungen eine Vorhersage getroffen. Für Faktoren, die sich nicht nur in eine Richtung auf die Selbsteinschätzung von allen Fällen auswirken, sondern in eine Richtung auf die Fälle, die ihre Leistung überschätzen und in die andere Richtung auf die Selbsteinschätzung von Fällen, die ihre Leistung unterschätzen, wird der Einfluss auf die Genauigkeit der Selbsteinschätzung auf diese Weise übersehen, falls der Faktor das Ausmaß an Über- und Unterschätzung gleichermaßen reduziert, weil sich diese statistisch gegenseitig ausgleichen. Aufgrund der allgemeinen Überschätzung der Deutschkompetenzen durch die Teilnehmenden kann jedoch die Annahme getroffen werden, dass für Faktoren, die die Genauigkeit von Selbsteinschätzungen beeinflussen, insgesamt die Reduktion der Überschätzung gegenüber der Reduktion der Unterschätzung überwiegt und deshalb in Summe ein Effekt auf die Höhe der Selbsteinschätzung gefunden wird. Dabei wird jedoch in Kauf genommen, dass der Einfluss von Faktoren auf die Selbsteinschätzungen in dem Ausmaß unterschätzt wird, in dem sie auch Unterschätzungen in die entgegengesetzte Richtung systematisch beeinflussen. Für die Fähigkeit zum schlussfolgernden Denken und die Teilnahme an einem Deutschtest bedeutet dies, dass Einflüsse

auf die Genauigkeit der Selbsteinschätzungen größer ausfallen könnten als die gefundenen Einflüsse auf die Selbsteinschätzungen selbst.

Aus folgenden Gründen wurde es dennoch bevorzugt, die Selbsteinschätzung statt eines Maßes der Genauigkeit der Selbsteinschätzung als zu erklärende Variable heranzuziehen: Eine Modellierung der Genauigkeit als Betrag der Differenz zwischen der Selbsteinschätzung und einer für das gemessene Kompetenzniveau als angemessen definierten Selbsteinschätzung oder als Betrag des Residuums für den Zusammenhang zwischen Selbsteinschätzung und objektiv gemessener Kompetenz hätte den Nachteil, dass diese Variable sehr empfindlich davon abhinge, wie definiert wird, welche Selbsteinschätzung bei welcher Kompetenz angemessen ist oder davon, wie die Teilnehmenden ihre Kompetenzen selbst einschätzen, falls man ein Residualmaß verwendet. Zudem konnten auf diese Weise alle betrachteten Einflussfaktoren in dasselbe Modell aufgenommen werden. So wurde jeweils der Einfluss der anderen Variablen kontrolliert.

Dem muss jedoch hinzugefügt werden, dass nicht alle Variablen im Modell enthalten sind, für die ein Einfluss auf die Selbsteinschätzungen vorhergesagt werden könnte. Neben den untersuchten Einflussfaktoren gibt es weitere Faktoren, für die ein Einfluss auf die Selbsteinschätzungen aus theoretischer Sicht zu erwarten wäre und untersucht werden sollte. Nicht berücksichtigt wurde in diesem Modell zum Beispiel der Einfluss externaler Referenzrahmen, wie er z.B. im BFLPE beschrieben ist. Um diesen Einfluss zu untersuchen, wären Daten geeignet, die die objektiven und subjektiven Kompetenzmaße von ganzen Schulklassen und idealerweise zusätzlich vom privaten Umfeld der Teilnehmenden umfassen. So könnte mithilfe einer Mehrebenenstruktur der Einfluss des Kompetenzniveaus des sozialen Umfelds auf die Selbsteinschätzungen untersucht werden. Weiterhin wurde der Dunning-Kruger-Effekt nicht berücksichtigt, also dass Personen mit niedrigerer Kompetenz eher zur Überschätzung dieser Kompetenz neigen als Personen mit höherer Kompetenz. Ergänzende Forschung, in der ein Gesamtmodell untersucht wird, in dem noch weitere mögliche Einflussfaktoren modelliert werden, wäre demzufolge wünschenswert.

5.3 Vergleich verschiedener Arten von Selbsteinschätzungsitems

Um herauszufinden, mit welcher Art von Selbsteinschätzungen die Deutschkompetenzen der jugendlichen Flüchtlinge am genauesten erfasst werden können, wurde zuletzt geprüft, ob bestimmte Arten von Selbsteinschätzungsitems bestimmte Vorteile haben. Beim Vergleich mit dem Standard-Selbsteinschätzungsitem wies das Schieberegler-Item entgegen der Vorhersage in Hypothese 3a keine geringere Schiefe in der Verteilung der Antworten auf und es konnte entgegen der Hypothese 3b auch nicht besser zwischen Personen mit unterschiedlichem Kompetenzniveau im Verstehen der deutschen Sprache diskriminieren. Die insgesamt breitere Verteilung des

Schieberegler-Items ging demnach nicht mit einer besseren Diskriminationsfähigkeit einher. Eine im Vergleich zur Standard-Skala bessere Diskriminationsfähigkeit der Schieberegler-Skala wurde vorhergesagt, weil die Standard-Skala nur wenige Antwortkategorien insbesondere im positiven Bereich umfasst und deshalb keine differenzierteren Einschätzungen zulässt (vgl. Kapitel 2.3.1). Eine höhere Anzahl an Antwortkategorien, wie es bei den Schieberegler-Items der Fall ist, hat jedoch nur einen Vorteil, wenn die Teilnehmenden auch entsprechend genau und einheitlich zwischen den Kategorien differenzieren (vgl. Krosnick & Presser, 2010; s. Abschnitt 1.7.2). Dies scheint nicht der Fall zu sein. Die Teilnehmenden scheinen keine so präzise und gleichzeitig einheitliche Unterscheidungen zwischen den Antwortkategorien der Endpunkt-gelabelten zehnstufigen Skala vorzunehmen, dass die höhere Anzahl an Antwortkategorien einen Vorteil gegenüber der kürzeren Standard-Skala hätte. Obwohl sich die Antworten der Teilnehmenden auf mehr Kategorien verteilen, habt das Schieberegler-Item gegenüber dem Standard-Item hier insgesamt keinen Vorteil.

Bei dem Vergleich-Item war der Deckeneffekt, der bei dem Standard-Item problematisiert wurde, nicht bzw. deutlich geringer vorhanden, was der Vorhersage in Hypothese 4a entspricht. Der Deckeneffekt konnte reduziert werden, weil eine Referenzgruppe mit durchschnittlich besseren Deutschkompetenzen als die der Teilnehmenden vorgegeben wurde. Trotzdem konnte das Vergleich-Item entgegen der Annahme von Hypothese 4b nicht signifikant besser zwischen den Deutschkompetenzen der Teilnehmenden diskriminieren, wozu die im Folgenden erläuterte Problematik beigetragen hat.

In Abschnitt 2.3.2 wurde bereits darauf hingewiesen, dass Muttersprachlerinnen und -sprachler deshalb keine besonders geeignete Referenzgruppe darstellen, weil sich Personen bevorzugt mit anderen vergleichen, deren Fähigkeiten auf einem ähnlichen Niveau liegen (Festinger, 1954), was in dem Fall nicht gegeben ist. Bei den Vergleich-Items ist der Unterschied zwischen den mit dem PPVT-4 gemessenen sprachlichen Kompetenzen der jugendlichen Flüchtlinge und der Normstichprobe gleichaltriger Jugendlicher, deren Muttersprache überwiegend Deutsch ist, sogar so groß, dass ein angemessener Vergleich kaum möglich ist. In *Abbildung 27* in Abschnitt 4.4.2 wird ersichtlich, dass die wenigsten Jugendlichen einen Summenwert im PPVT-4 erzielt haben, der besser ist als der Wert, der zwei Standardabweichungen unter dem Mittelwert der Normstichprobe liegt. Die Definition in Abschnitt 4.4.2 sieht jedoch vor, dass nur Personen, deren Summenwert weniger als zwei Standardabweichungen vom Mittelwert der Normstichprobe abweicht, ihre Kompetenz als *fast genauso gut* oder *genauso gut wie ein Deutscher* einstufen sollten, wobei diese Definition tendenziell großzügig gehalten ist. Auch den unterdurchschnittlichen Bereich mit einer Abweichung von ein bis zwei Standardabweichungen könnte man alternativ bereits als *schlechter als ein Deutscher* definieren. Für die allermeisten jugendlichen Flüchtlinge wären also die Kategorien *schlechter als ein Deutscher* oder *viel schlechter als ein Deutscher* angemessener. Da in dem Bereich, auf dem sich die Summenwerte der Stichprobe der jugendlichen Flüchtlinge verteilen, kaum

Summenwerte der Muttersprachler liegen, gibt es für diesen Bereich jedoch keine Normwerte, sodass die genauere Definition der schlechten Kategorien im Vergleich zur Normstichprobe nicht weiter ausdifferenziert werden kann. Insgesamt macht also die sehr geringe Überlappung der Verteilungen der Summenwerte der jugendlichen Flüchtlinge und der Muttersprachler eine angemessene Zuordnung der Kompetenzniveaus der jugendlichen Flüchtlinge im Vergleich zu den Muttersprachlern fast unmöglich und der Kompetenzbereich der jugendlichen Flüchtlinge wird durch die Skala nicht differenziert genug abgebildet. Aufgrund dieser Tatsachen scheinen gleichaltrige Muttersprachler keine angemessene Referenzgruppe für die Selbsteinschätzungen der sprachlichen Kompetenzen im Deutschen für jugendliche Flüchtlinge, die erst seit durchschnittlich ca. zweieinhalb Jahren in Deutschland leben, zu sein. Insgesamt konnten die Deutschkompetenzen mit den Vergleich-Items demnach nicht genauer erfasst werden als mit den Standard-Items.

Auch die Can-Do-Statements zum Verstehen und Sprechen korrelierten zu beiden Messzeitpunkten nicht signifikant stärker mit den objektiven Kompetenzmaßen als das Standard-Item zum Verstehen der deutschen Sprache, weshalb Hypothese 5a verworfen werden musste. Jedoch scheinen die Can-Do-Statements Veränderungen der sprachlichen Kompetenz im Deutschen zwischen zwei Messzeitpunkten besser wiederzugeben als die Standard-Items zum Verstehen und Sprechen der deutschen Sprache, was Hypothese 5b bestätigte.

Für die Can-Do-Statements wurde erwartet, dass diese besser zwischen den Kompetenzen der Teilnehmenden diskriminieren, weil sie eindeutiger formuliert sind, indem sie sich auf alltägliche Situationen und weniger umfangreiche Fähigkeiten beziehen und weniger Interpretationsspielraum lassen als z.B. die Standard-Items. Für spezifischere Items mit vertrauterem Inhalt wurden auch in anderen Studien höhere Korrelationen mit objektiven Kompetenzmaßen gefunden (Freund & Kasten, 2012; LeBlanc & Painchaud, 1985; vgl. Abschnitt 1.7.2). Die Teilnehmenden müssen weniger Informationen integrieren, was die Schwierigkeit der Beantwortung der Selbsteinschätzungsfrage reduziert. Dadurch sollte die Motivation und die Gründlichkeit bei der Beantwortung erhöht werden (Krosnick & Presser, 2010; vgl. Abschnitt 1.7.1). Diese Annahmen wurden jedoch nicht bestätigt. Dass die Can-Do-Statements nicht erwartungsgemäß stärker mit den objektiven Kompetenzmaßen korrelieren als die Standard-Items kann verschiedene Gründe haben. Es könnte zum einen daran liegen, dass die Statements teilweise relativ lang sind und insbesondere Teilnehmende, die den Fragebogen nicht in ihrer Muttersprache beantwortet haben, diese möglicherweise nicht richtig verstanden haben. Dieses Argument würde auch zu der Erklärung beitragen, warum in der siebten Erhebungswelle für die Can-Do-Statements eine höhere Korrelation mit den objektiven Kompetenzmaßen gefunden wurde als in der ersten Erhebungswelle, denn zwischen der ersten und der siebten Erhebungswelle haben die Teilnehmenden ihre Deutschkompetenzen verbessert und Teilnehmende, die den Fragebogen auf Deutsch beantwortet haben, haben die Statements möglicherweise besser verstanden als zum Zeitpunkt der ersten Erhebungswelle.

Eine andere Möglichkeit ist, dass die Teilnehmenden die Statements unterschiedlich interpretiert haben bzw. implizit von unterschiedlichen Grenzen ausgegangen sind, ab wann sie einem Statement zugestimmt haben. So bieten auch die Can-Do-Statements entgegen der Intention einen Interpretationsspielraum. Beispielsweise ist für das Item „einfache Gespräche über vertraute Themen führen“ nicht genau definiert, wie einfach die Gespräche sein müssen und wie fehlerfrei die Kommunikation ablaufen sollte. In der Studie von LeBlanc und Painchaud (1985) sind die Situationen, auf die sich die Statements beziehen, ausführlicher beschrieben, wodurch die Vertrautheit und Eindeutigkeit noch höher ist. Dadurch konnten sie sehr hohe Korrelationen mit den objektiven Maßen erreichen und diese deutlich verbessern gegenüber weniger ausführlichen Can-Do-Statements.

Weiterhin weist der Summenwert der Can-Do-Statements einen deutlichen Deckeneffekt auf und differenziert deshalb insbesondere zwischen denjenigen Teilnehmenden schlecht, die sich auf Deutsch viel zutrauen. Ein schwierigeres Item zum Verstehen und Sprechen musste von der Summenwertbildung ausgeschlossen werden, da die Skala sonst nicht eindimensional gewesen wäre. Möglicherweise könnte die Korrelation des Summenwerts der Can-Do-Statements mit den objektiven Kompetenzmaßen jedoch erhöht werden, wenn zusätzliche schwierigere Statements ergänzt würden, die dieselbe Dimension messen wie die bereits vorhandenen Statements.

Weiterhin wurde für die Can-Do-Statements vorhergesagt, dass diese die Veränderung der Kompetenz zwischen zwei Erhebungszeitpunkten besser abbilden können als die Standard-Items, was anhand der Analyseergebnisse bestätigt wurde. Dies ist mit hoher Wahrscheinlichkeit darauf zurückzuführen, dass die Can-Do-Statements weniger von Referenzrahmen abhängen, welche sich ebenfalls mit der Zeit verändern, als die allgemein formulierten Standard-Items und somit über die Zeit hinweg konsistenter interpretiert werden.

Insgesamt haben die Can-Do-Statements also nicht besser zwischen den Deutschkompetenzen der Teilnehmenden diskriminiert als das Standard-Item. Dies könnte jedoch u.a. auf den Deckeneffekt zurückzuführen sein und durch die Hinzunahme schwierigerer Items könnte die Genauigkeit der Can-Do-Statements verbessert werden. Die Veränderung zwischen zwei Messzeitpunkten wurde mit den Can-Do-Statements besser erfasst als mit den Standard-Items, was einen Vorteil der Can-Do-Statements darstellt.

5.4 Allgemeine Limitationen

Auf einzelne spezifische Limitationen der Analysen wurde in den vorangegangenen Kapiteln 5.1 bis 5.3 eingegangen. Übergeordnete Limitationen, die mehrere Analysen betreffen, werden in den folgenden Abschnitten diskutiert. Dabei wird zuerst auf die Selbsteinschätzungen

eingegangen und anschließend auf die objektiven Kompetenzmaße. Abschließend wird die Selektivität der Stichproben diskutiert.

5.4.1 Selbsteinschätzungen

Hinsichtlich der Selbsteinschätzungen ist zunächst hervorzuheben, dass es ein besonderer Vorteil der ReGES-Studie ist, dass vier verschiedene Arten von Selbsteinschätzungsitems aufgenommen wurden, was den direkten Vergleich zwischen diesen ermöglicht hat. Dabei wurden die im Folgenden ausgeführten kleineren Einschränkungen in Kauf genommen.

Zum einen musste aufgrund der begrenzten Befragungszeit das Selbsteinschätzungsmodul zu den Deutschkompetenzen möglichst kurzgehalten werden. Aus diesem Grund wurden allen Teilnehmenden entweder die Vergleich- oder die Schieberegler-Items vorgegeben. Deshalb sind die Stichproben und damit auch die Teststärken für diese Items kleiner und es konnten nicht alle Vergleiche anhand derselben Stichprobe durchgeführt werden. Da die Vorgabe der Items jedoch zufällig erfolgte, ist davon auszugehen, dass sich die beiden Stichproben nicht systematisch unterscheiden. Außerdem war auch die Hälfte der Stichprobe so groß, dass eine gute Teststärke erzielt werden konnte.

Weiterhin wurden die Vergleich- und die Schieberegler-Items nur in der ersten Erhebungswelle, nicht jedoch in der siebten Erhebungswelle eingesetzt. Es konnte deshalb nicht geprüft werden, ob die Hypothesen zu dem späteren Messzeitpunkt bestätigt worden wären, wie es bei den Can-Do-Statements der Fall war. Möglicherweise hätten z.B. die zum zweiten Messzeitpunkt ca. zwei Jahre älteren Teilnehmenden präzisere Unterscheidungen der Skalenpunkte der Schieberegler-Items vorgenommen.

Zudem wurden die Items immer in derselben Reihenfolge im Fragebogen vorgegeben. Transfereffekte zwischen den Items sind also möglich. Möglicherweise haben Personen, die die Standarditems mit *sehr gut* beantwortet haben, z.B. bei den Can-Do-Statements eher dazu tendiert, möglichst viele Statements auszuwählen. Um diese Effekte kontrollieren zu können, wäre eine randomisierte Reihenfolge der Items von Vorteil gewesen.

5.4.2 Objektive Messung der sprachlichen Kompetenzen im Deutschen

Für die verschiedenen durchgeführten Analysen wurden objektive Maße der Sprachkompetenzen als Kriterien herangezogen. Dazu wurde der rezeptive Wortschatz mit dem PPVT-4 und das Grammatikverständnis mit dem TROG-D erfasst. Damit wurden hohe wissenschaftliche Standards erfüllt. Bei den genannten Tests handelt es sich um etablierte hochstandardisierte Verfahren zur Sprachkompetenzmessung, die eine Vielzahl an psychometrischen Qualitätskriterien erfüllen. Zudem wurden die Instruktionen für den Einsatz in der Flüchtlingsstichprobe in drei weitere

Sprachen übersetzt, um sicherzustellen, dass die Teilnehmenden die Aufgaben verstehen. Außerdem wurden die Tests computerbasiert durchgeführt, was die Standardisierung noch weiter erhöht hat. Neben diesen Vorteilen der durchgeführten Kompetenztestungen bleiben zwei Limitationen zu diskutieren.

Zum einen erfassen die Tests ausschließlich das Hörverstehen und decken somit nicht den gesamten sprachlichen Kompetenzbereich ab, wie z.B. Fähigkeiten im Sprechen, welche auch die Aussprache und produktive Grammatik umfassen, oder im Lesen oder Schreiben. Deshalb wurden in dieser Arbeit hauptsächlich die Selbsteinschätzungen zum Verstehen der deutschen Sprache betrachtet, teilweise wurden jedoch auch die Selbsteinschätzungen zum Sprechen der deutschen Sprache miteinbezogen. Obwohl die verschiedenen sprachlichen Kompetenzbereiche korrelieren (z.B. Lenhard et al., 2015), könnte das Ausmaß, in dem die Selbsteinschätzungen die tatsächlichen zugrundeliegenden Kompetenzen messen, leicht unterschätzt werden, weil die Selbsteinschätzungen zusätzliche Kompetenzbereiche erfassen, die die Kompetenztests nicht erfassen. Um die sprachlichen Kompetenzen in zukünftigen Studien umfassender mit objektiven Testverfahren zu messen, könnten weitere standardisierte Testverfahren eingesetzt werden, die auch produktive Kompetenzen im Sprechen und Schreiben sowie Lesekompetenzen erfassen. Dann könnten auch die hier nicht näher betrachteten Selbsteinschätzungen zur Lese- und Schreibkompetenz untersucht werden.

Zum anderen stellt sich die Frage, ob die adaptive Gestaltung des PPVT-4 in der Flüchtlingsstichprobe wie vorgesehen funktioniert und insbesondere die Annahmen, dass Items oberhalb des Deckensets größtenteils nicht gelöst worden wären und Items unterhalb des Bodensets größtenteils gelöst worden wären, auch in dieser Stichprobe erfüllt ist (vgl. Obry et al., 2021). Das Abbruchkriterium im PPVT-4 hat die Funktion, dass nicht alle Teilnehmenden alle 228 Items bearbeiten müssen, sondern diese nur Items in dem Schwierigkeitsbereich bearbeiten, der ungefähr ihrer Fähigkeit entspricht. Im PPVT-4 sind die Items in aufsteigender Schwierigkeit angeordnet und der Test wird abgebrochen, sobald das Boden- und das Deckenset identifiziert wurden, also das niedrigste Itemset, in dem keine bzw. fast keine Fehler mehr gemacht wurden und das höchste Itemset, in dem der Anteil richtiger Antworten nur noch auf Zufallsniveau liegt (vgl. Lenhard et al., 2015). Ist der Anstieg der Itemschwierigkeit über die Sets gewährleistet, könnten die Teilnehmenden fast alle Items unterhalb des Bodensets beantworten und müssten bei fast allen Items oberhalb des Deckensets raten, weil sie die Antworten nicht kennen. In diesem Fall ist es vertretbar, die betroffenen Items zur Einsparung von Ressourcen und um eine Demotivation der Teilnehmenden zu verhindern, nicht durchzuführen und als richtig bzw. falsch gelöst zu kodieren. Ist der Anstieg der Itemschwierigkeit über die Sets jedoch nicht gegeben, könnte es sein, dass das Bodenset erreicht wird, aber dennoch eine signifikante Anzahl an Items unterhalb des Bodensets nicht gewusst worden wäre, die dann fälschlich als richtig gelöst kodiert werden. Auf der anderen Seite könnte das

Deckenset erreicht werden, weil ein bestimmtes Itemset für die oder den Teilnehmenden besonders schwierig war, diese Person hätte aber möglicherweise eine bedeutsame Menge an Items oberhalb des Deckensets richtig beantworten können. Die Kodierung dieser Items als falsch gelöst würde die Aussagekraft und Vergleichbarkeit der Testergebnisse beeinträchtigen. Bei der Untersuchung der PPVT-4-Daten in der Stichprobe der jugendlichen Flüchtlinge hat sich gezeigt, dass die Varianz der Itemschwierigkeiten innerhalb der Sets sehr groß ist und diese zwischen den Sets stark überlappen (Obry et al., 2021). Weil außerdem die jugendliche Flüchtlinge beim Zweitspracherwerb die Vokabeln in einer anderen Reihenfolge erlernen, als es beim Erstspracherwerb der Fall ist (Appel, 1996), kann angenommen werden, dass die Testlogik des PPVT-4 in dieser Stichprobe weniger gut funktioniert als bei muttersprachlichen Teilnehmenden. Die Konsequenz ist, dass die Testergebnisse wahrscheinlich etwas weniger verlässlich und vergleichbar sind als die Testergebnisse muttersprachlicher Stichproben. Diese Annahme sollte in ergänzenden Analysen näher untersucht werden. Dennoch waren die Ergebnisse der Qualitätsprüfung der Skala überwiegend zufriedenstellend und der PPVT-4 erfüllt hohe Standards bei der Messung des rezeptiven Wortschatzes der jugendlichen Flüchtlinge.

5.4.3 Selektivität der Stichprobe

In Abschnitt 3.1.4 wurde die Selektivität der Analysestichproben untersucht, indem die Stichproben jeweils mit den gültigen Fällen der ReGES-Stichprobe zum ersten Erhebungszeitpunkt verglichen wurden, die nicht Teil der Analysestichproben waren. Mögliche Auswirkungen im Folgenden zusammengefasster Ergebnisse der Selektivitätsanalyse auf die Ergebnisse dieser Studie werden in diesem Abschnitt diskutiert. Für die Deutschkompetenzen der Teilnehmenden wurde keine Selektivität nachgewiesen, sie konnte jedoch aufgrund systematisch fehlender Deutschkompetenzwerte auch nicht ausgeschlossen werden. Falls es einen Unterschied hinsichtlich der Deutschkompetenz gab, wäre zu erwarten, dass die Deutschkompetenzen der Analysestichproben durchschnittlich in geringem Umfang besser waren als die der Stichproben ausgeschlossener Fälle, da Personen, die weder Arabisch noch Kurmandschi noch Englisch verstanden, nur befragt und getestet wurden, wenn sie dazu ausreichende Deutschkompetenzen besaßen. Die Selbsteinschätzungen der Deutschkompetenzen der Analysestichproben waren im Durchschnitt jedoch gleich oder niedriger als die Selbsteinschätzungen der Deutschkompetenzen der Stichproben ausgeschlossener Fälle. Hinsichtlich der kognitiven Grundfähigkeiten kann die Selektivität der Analysestichprobe 1 aufgrund fehlender Kompetenzwerte in der Stichprobe der ausgeschlossenen Fälle nicht beurteilt werden. In der Analysestichprobe 2 gab es kleine bis mittlere Unterschiede hinsichtlich der kognitiven Grundfähigkeiten zugunsten der Analysestichprobe. Weiterhin war der Anteil an Personen, die in einer privaten Unterkunft und nicht in einer Gemeinschaftsunterkunft wohnten in den

Analysestichproben höher als in den Stichproben ausgeschlossener Fälle. Hinsichtlich der Herkunft war insbesondere der Anteil von Personen aus Syrien in den Analysestichproben größer und der Anteil von Personen aus Afghanistan kleiner als in den Stichproben ausgeschlossener Fälle.

Für den Fall, dass die Analysestichproben hinsichtlich der Deutschkompetenz in geringem Umfang selektiv waren, wäre dies hauptsächlich bei den Analysen zur allgemeinen Verzerrung der Selbsteinschätzungen zu berücksichtigen. Hier wurde für die Beurteilung der Genauigkeit der Selbsteinschätzungen definiert, dass die Skalen der Standard- und der Schieberegler-Items der Verteilung der gleichaltrigen jugendlichen Flüchtlinge entsprechen sollen (vgl. Abschnitt 3.3.2.3). Hätten die gleichaltrigen jugendlichen Flüchtlinge tatsächlich durchschnittlich etwas schlechtere Deutschkompetenzen, wäre die Definition einer genauen Selbsteinschätzung an diese Verteilung anzupassen und die Überschätzung würde im Durchschnitt etwas geringer ausfallen, weil die Deutschkompetenzen der untersuchten Stichproben tatsächlich etwas besser wären als die der gleichaltrigen jugendlichen Flüchtlinge in Deutschland insgesamt. Da wenn überhaupt geringe Unterschiede in der durchschnittlichen Deutschkompetenz der Stichproben zu erwarten sind, würde sich dies auch nur in geringem Umfang auf die Ergebnisse auswirken und es ist weiterhin davon auszugehen, dass die jugendlichen Flüchtlinge ihre Deutschkompetenzen auch bei den Standard- und Schieberegler-Items überschätzen.

Da sich Fähigkeiten zum schlussfolgernden Denken negativ auf die Höhe der Selbsteinschätzungen auswirken (vgl. Kapitel 4.5), und zumindest die Analysestichprobe 2 leicht überdurchschnittlich im Test dieser Fähigkeiten abgeschnitten hat, könnten diese ihre Deutschkompetenzen etwas weniger überschätzt haben als es für jugendliche Flüchtlinge in Deutschland insgesamt der Fall wäre. Aufgrund der geringen Effektstärken ist jedoch auch hier nur ein geringer Unterschied zu erwarten, der die Ergebnisse dieser Arbeit nicht wesentlich beeinflusst.

Hinsichtlich der Unterbringung lassen sich folgende Auswirkungen der Selektivität der Analysestichproben auf die Ergebnisse dieser Arbeit vermuten. Aufgrund von Referenzrahmeneffekten wie dem BFLPE (s. Abschnitt 1.2.4) könnten Teilnehmende, die in einer Gemeinschaftsunterkunft untergebracht waren, bei gleicher Deutschkompetenz ihre Deutschkompetenz höher eingeschätzt haben als Teilnehmende, die privat untergebracht und möglicherweise vermehrt von Personen mit besseren Deutschkompetenzen umgeben waren. Da der Anteil von Teilnehmenden, die in einer privaten Unterkunft untergebracht waren, in den Analysestichproben überdurchschnittlich hoch waren, ist es wahrscheinlich, dass die Teilnehmenden ihre Deutschkompetenzen vergleichsweise niedriger eingeschätzt haben als es bei einer in dieser Hinsicht nicht-selektiven Stichprobe zu erwarten wäre. Auch hier sind jedoch nur geringe Effekte und damit Auswirkungen auf die Ergebnisse dieser Arbeit anzunehmen.

Zusammenfassend könnten sich mögliche Selektivitätseffekte hauptsächlich darauf ausgewirkt haben, wie die Überschätzung der Deutschkompetenzen durch die Teilnehmenden in dieser

Arbeit beurteilt wurde. Aufgrund der möglicherweise leicht überdurchschnittlichen Fähigkeiten der Analysestichproben zum schlussfolgernden Denken und dem größeren Anteil an Personen, die in einer privaten Unterkunft untergebracht waren, könnte das Ausmaß der Überschätzung der Deutschkompetenzen in einer nicht-selektiven Stichprobe etwas größer sein, als es hier gefunden wurde.

Für die Generalisierbarkeit der Ergebnisse könnte insbesondere die Selektivität hinsichtlich der Herkunftsländer eine Rolle spielen. In den Analysestichproben überwogen deutlich Personen mit syrischer Herkunft. Ergebnisse dieser Arbeit sind also insbesondere für jugendliche Geflüchtete aus Syrien gültig und nicht auf sämtliche Herkunftsgruppen ohne weiteres generalisierbar.

5.5 Allgemeine Implikationen

Aus dem in dieser Arbeit gewonnenen Wissen können Implikationen für die Interpretation von Forschungsergebnissen zu Deutschkompetenzen jugendlicher Flüchtlinge abgeleitet werden, für die Deutschkompetenzen mittels Selbsteinschätzungen gemessen wurden. Darauf wird im folgenden Abschnitt 5.5.1 eingegangen. Weiterhin können Möglichkeiten zur Optimierung der Selbsteinschätzungsitems abgeleitet werden, worauf in Abschnitt 5.5.2 eingegangen wird.

5.5.1 Konsequenzen der Messung von Deutschkompetenzen mittels Selbsteinschätzungen

Für den Fall, dass man sprachliche Kompetenzen im Deutschen jugendlicher Flüchtlinge ausschließlich mit Selbsteinschätzungen misst, muss man auf Grundlage der hier gefundenen Ergebnisse davon ausgehen, dass die Rangordnung der Teilnehmenden nur mittelmäßig mit der tatsächlichen Rangordnung der Kompetenzen der Teilnehmenden übereinstimmt. Vergleicht man zur Veranschaulichung dieses Ergebnisses die Verteilungen der objektiven Kompetenzwerte der Gruppen von Teilnehmenden, die jeweils eine bestimmte Antwortkategorie gewählt haben (s. *Abbildung 23* und *Abbildung 24* in Abschnitt 4.4.1), sind zwar die Mittelwerte der objektiven Kompetenzen dieser Gruppen entsprechend deren Selbsteinschätzung geordnet, die Verteilungen selbst überlappen sich jedoch stark. Es gibt also z.B. eine große Gruppe Teilnehmender, die ihre Deutschkompetenzen als *sehr gut* einschätzen, aber im Vergleich mit einem großen Anteil an Teilnehmenden, die ihre Deutschkompetenzen als *eher gut* einschätzen, schlechtere Deutschkompetenzwerte erzielt haben.

Weiterhin können die Selbsteinschätzungen nicht dazu herangezogen werden, um das tatsächliche Sprachkompetenzniveau der jugendlichen Flüchtlinge anzugeben. Zum einen muss man davon ausgehen, dass die Teilnehmenden ihre Deutschkompetenzen deutlich überschätzen und

somit auch das Sprachkompetenzniveau der Gruppe der jugendlichen Flüchtlinge auf diese Weise überschätzt würde. Beispielsweise haben nach den Ergebnissen der Selbsteinschätzungen mit den Vergleich-Items bereits über die Hälfte der jugendlichen Flüchtlinge nach einem ca. zweieinhalbjährigen Aufenthalt in Deutschland fast ein muttersprachliches Niveau der deutschen Sprachkompetenz erreicht. Der Vergleich der Ergebnisse des PPVT-4 der jugendlichen Flüchtlinge mit der Normstichprobe hat jedoch gezeigt, dass erst ein sehr geringer Anteil der jugendlichen Flüchtlinge sich im Bereich des muttersprachlichen Niveaus bewegt. Ausschließlich die Ergebnisse dieser Selbsteinschätzungen als Maß für das Sprachniveau der jugendlichen Flüchtlinge heranzuziehen, würde demnach zu falschen Aussagen und einer Überschätzung des Sprachniveaus führen. Zum anderen ist es anhand von Items, die keinen Referenzrahmen vorgeben, wie z.B. die Standard- und die Schieberegler-Items, grundsätzlich nicht möglich, Aussagen zum absoluten Sprachniveau der Teilnehmenden zu treffen, da die Antworten auch davon abhängen, wie die Teilnehmenden die Skalenpunkte interpretieren. Interpretieren die Teilnehmenden die Skalenpunkte wie angenommen so, dass diese die Verteilung anderer Geflüchteter in ihrem direkten Umfeld abbilden, sagen die Selbsteinschätzungen nur etwas über das Kompetenzniveau im Vergleich zu dieser Gruppe aus und sind nicht hilfreich, um absolutere Informationen über das allgemeine Sprachniveau der Gruppe zu erhalten, sofern das Sprachniveau dieser Gruppe nicht bereits bekannt ist. Die Can-Do-Statements lassen hingegen etwas präzisere Aussagen zum Sprachniveau der Gruppe zu, da konkretere Aussagen zu den Sprachkompetenzen getroffen werden können, wie z.B., dass ein bestimmter Anteil der jugendlichen Flüchtlinge angibt, einfache Gespräche über vertraute Themen führen zu können und sich diese Statements wiederum auf den Kompetenzstufen des GER einordnen lassen. Jedoch ist auch hier nicht sichergestellt, dass die Jugendlichen ihre Kompetenzen nicht überschätzen und auch diese Statements scheinen nicht einheitlich verstanden und interpretiert zu werden.

Möchte man die Selbsteinschätzungen der Sprachkompetenzen der jugendlichen Flüchtlinge mit anderen Variablen in Beziehung setzen, um z.B. die Determinanten der Sprachkompetenzen zu bestimmen, muss man zudem beachten, dass die Selbsteinschätzungen systematisch von bestimmten Faktoren beeinflusst werden, die auch mit den anderen Variablen zusammenhängen und so die Ergebnisse verzerren können. Zum Beispiel könnte man den Einfluss kognitiver Fähigkeiten auf die Sprachkompetenzen der jugendlichen Flüchtlinge untersuchen wollen. Misst man die Sprachkompetenzen jedoch mittels Selbsteinschätzung, würde ein positiver Einfluss der kognitiven Fähigkeiten unterschätzt werden, da sich diese unabhängig von der tatsächlichen Sprachkompetenz negativ auf die Höhe der Selbsteinschätzungen auswirken. Ebenso könnte der Einfluss des Engagements beim Deutschlernen auf die Deutschkompetenzen auf diese Weise überschätzt werden, da Personen, die viel Engagement beim Deutschlernen zeigen, dazu neigen, ihre Deutschkompetenzen zu überschätzen. Ein möglicher positiver Zusammenhang zwischen dem

Engagement beim Deutschlernen könnte also verstärkt werden, wenn man Selbsteinschätzungen statt objektiver Kompetenzmaße verwendet. Neben den in dieser Arbeit identifizierten Faktoren, die die Selbsteinschätzungen beeinflussen, ist auf Grundlage der Literatur von weiteren Faktoren auszugehen, die die Selbsteinschätzungen systematisch beeinflussen. Der Einfluss mancher dieser Faktoren konnte in dieser Arbeit nicht untersucht werden und der Einfluss anderer Faktoren wurde nicht bestätigt. Weitere Analysen bzw. Replikationen müssen zeigen, ob dies auf methodische Probleme zurückzuführen ist oder ob bestimmte Faktoren die Selbsteinschätzungen der Deutschkompetenzen jugendlicher Flüchtlinge nicht wie erwartet beeinflussen, die Selbsteinschätzungen anderer Stichproben jedoch schon, oder ob sich diese Faktoren tatsächlich nicht wie erwartet auf die Selbsteinschätzungen auswirken.

5.5.2 Möglichkeiten zur Optimierung der Selbsteinschätzungen

Im Rahmen dieser Arbeit wurden verschiedene Arten von Selbsteinschätzungsitems miteinander verglichen, um festzustellen, ob die sprachlichen Kompetenzen der jugendlichen Flüchtlinge im Deutschen mit bestimmten Items genauer erfasst werden können als mit anderen Items. Im Vergleich des Schieberegler-Items und des Vergleich-Items konnte als einziger Vorteil gegenüber dem Standard-Item eine Reduktion des Deckeneffekts gefunden werden. Die Can-Do-Statements haben die Kompetenzzuwächse zwischen den beiden Messzeitpunkten besser erfasst als das Standard-Item, sind jedoch umfangreicher und beanspruchen daher mehr Beantwortungszeit. Aus den Ergebnissen dieser Arbeit lassen sich zudem Möglichkeiten ableiten, wie die Selbsteinschätzungsitems verbessert werden könnten, um die Genauigkeit der Selbsteinschätzungen zu erhöhen.

In Bezug auf das Standard-Item hat sich gezeigt, dass sich die Antworten größtenteils auf die oberen beiden positiven Kategorien verteilen. Bei dem Schieberegler-Item mit einer zehnstufigen Skala wurde zwar erreicht, dass sich die Antworten auf mehr Kategorien verteilen, dadurch hat sich die Korrelation mit den objektiven Maßen jedoch nicht verbessert, vermutlich weil die Teilnehmenden die nur Endpunkt-gelabelten Kategorien weniger konsistent interpretiert haben. Deshalb stellt sich die Frage, ob es eine andere Möglichkeit gibt, die Anzahl der Kategorien, auf die sich die Antworten verteilen, zu erhöhen und gleichzeitig eine bessere Diskrimination zwischen den Kompetenzen der Teilnehmenden zu erreichen. Bei den Items der IAB-BAMF-SOEP-Befragung von Geflüchteten wurden z.B. zwei positive (*sehr gut* und *gut*), eine mittlere (*es geht*) und zwei schlechte (*eher schlecht* und *gar nicht*) Kategorien angeboten, während die Standard-Items der ReGES-Studie neben zwei positiven (*sehr gut* und *eher gut*) nur drei negative Antwortkategorien umfassten (*eher schlecht*, *sehr schlecht* und *gar nicht*). In der IAB-BAMF-SOEP-Befragung von Geflüchteten verteilten sich die Antworten der Teilnehmenden tatsächlich gleichmäßiger über die Antwortkategorien (s. Kapitel 5.1) als in der ReGES-Studie und ein erheblicher Anteil entfiel auf die mittlere

Antwortkategorie *es geht* und der Deckeneffekt war geringer (Niehues et al., 2021). Entsprechend wäre es lohnend zu untersuchen, ob bei jugendlichen Geflüchteten die Variation erhöht werden kann, wenn man die Labels der Standard-Items so ändert, dass sie neben den beiden positiven Kategorien noch eine mittlere Kategorie umfassen und dafür die Kategorie *sehr schlecht* oder die Kategorie *gar nicht* entfernt wird oder alternativ noch eine mittlere Kategorie hinzugefügt wird, sodass die Skala insgesamt aus sechs Antwortkategorien besteht. Falls sich die Teilnehmenden der ReGES-Studie im Zweifel mangels einer mittleren Kategorie für eine positive Kategorie entschieden haben, könnte so auch die allgemeine Überschätzung reduziert werden. Da die Differenzierung zwischen den Teilnehmenden durch die sehr geringe Variation der Selbsteinschätzungen eingeschränkt sein kann, könnte bei einer gleichmäßigeren Verteilung über die Antwortkategorien und konsistenter Interpretation der Antwortkategorien die Diskrimination der Items verbessert werden.

Dass die Vergleich-Items keine genaueren und aussagekräftigeren Selbsteinschätzungen ergeben, hängt wahrscheinlich damit zusammen, dass Gleichaltrige mit deutscher Muttersprache aus den in Kapitel 5.3 erläuterten Gründen keine geeignete Vergleichsgruppe darstellen. Deshalb ist das in anderen Studien gefundene Ergebnis, dass Personen ihre Kompetenzen genauer einschätzen können, wenn ein Referenzrahmen vorgegeben wird, nicht zu verwerfen. Die Vorgabe eines Referenzrahmens könnte dann hilfreich sein, wenn es eine geeignete Referenzgruppe gibt, die für den Vergleich herangezogen werden kann. In einer Stichprobe, wie sie in der ReGES-Studie zusammengesetzt wurde, ist es schwierig, eine solche geeignete Referenzgruppe zu finden, da wie in Abschnitt 2.3.2 erläutert z.B. auch das Sprachniveau anderer Zweitsprachlernerinnen und -lerner im direkten Umfeld der jugendlichen Flüchtlinge stark variieren kann. Eine geeignete Referenzgruppe könnte dann gefunden werden, wenn z.B. immer mehrere Teilnehmende eines Sprachkurses ihre Kompetenz relativ zu den anderen Teilnehmenden des Kurses einschätzen. Zumindest innerhalb dieser Gruppen könnten solche Vergleich-Items dann eine bessere Diskrimination zwischen den Teilnehmenden erzielen als Items ohne Vorgabe einer Referenzgruppe.

Etwas besser haben die Can-Do-Statements im Vergleich zu den Standard-Items abgeschnitten, aber auch für diese gibt es Möglichkeiten zur Optimierung. Insgesamt haben die Can-Do-Statements den Vorteil, dass man durch die Summenwertbildung mehr *Kategorien* erhält, ohne die Anforderungen an die Teilnehmenden hinsichtlich der Differenzierung zwischen den Antwortkategorien zu erhöhen. Ein Summenwert aus sieben Items resultiert in einer Variablen mit acht Ausprägungen, ohne dass die Teilnehmenden konsistent zwischen acht Antwortkategorien differenzieren müssen. Dennoch stellen auch die Can-Do-Statements Anforderungen an die Teilnehmenden hinsichtlich einer konsistenten Interpretation. Trotzdem sind die Items alltagsnäher und beschreiben konkretere Situationen, weshalb es den Teilnehmenden leichter fallen könnte, die Items zu beantworten. Zudem bieten die Statements den Vorteil, dass sie auch von Forschenden besser

zu interpretieren sind und in die Kompetenzniveaus des GER von A1 bis C2 eingeordnet werden können. Von den in dieser Arbeit verwendeten Items zum Verstehen und Sprechen der deutschen Sprache wurden jedoch viele Items von den meisten Teilnehmenden ausgewählt, sodass es einen Deckeneffekt gab. Dem könnte man entgegenwirken, indem man weitere oder andere Items für die Can-Do-Statements auswählt, z.B. aus dem Itempool des in Abschnitt 2.3.3 beschriebenen XS-Tests (Deutscher Volkshochschul-Verband e.V., 2011). Der Deckeneffekt könnte durch die Aufnahme schwierigerer Items reduziert werden und durch eine Erhöhung der Itemanzahl könnte auch die Variation vergrößert werden. Weitere Items müssten jedoch zunächst evaluiert werden. Um die Möglichkeiten zur Erfassung von sprachlichen Kompetenzen im Deutschen und ggf. auch in anderen Sprachen zu optimieren, wäre es entsprechend hilfreich, eine ausführliche Skala mit einer größeren Anzahl an Can-Do-Statements sowie verschiedene Kurzskalen an einer großen Stichprobe zu evaluieren. Zudem müsste auch für die Can-Do-Statements untersucht werden, inwiefern die Teilnehmenden ihre Sprachkompetenzen überschätzen. Dies wäre durch den Vergleich mit Testergebnissen, die ebenfalls eine Einstufung in die Kompetenzniveaus des GER vornehmen, möglich. Ein Nachteil der Can-Do-Statements ist, dass diese in der Regel mehr Items umfassen als die einzelnen Selbsteinschätzungsitems mit Likert-Skala und deshalb mehr Befragungszeit beanspruchen.

5.6 Fazit

Aus den Ergebnissen dieser Studie kann geschlussfolgert werden, dass Selbsteinschätzungen jugendlicher Flüchtlinge der Kompetenzen zum Verstehen der deutschen Sprache nur mittelmäßig zwischen deren tatsächlichen Deutschkompetenzen differenzieren und dass die jugendlichen Flüchtlinge ihre Deutschkompetenzen durchschnittlich überschätzen. Auch die Variation der tatsächlichen Kompetenzen wird nicht angemessen abgebildet. Zudem werden sie systematisch von Faktoren wie der Fähigkeit zum schlussfolgernden Denken und dem Engagement beim Deutschlernen beeinflusst. Deshalb ist die Verwendung von objektiven Testverfahren zur Erfassung der Deutschkompetenzen jugendlicher Flüchtlinge grundsätzlich zu bevorzugen. Falls Selbsteinschätzungen dennoch als ökonomischere Alternative verwendet werden, sind die Ergebnisse mit entsprechender Vorsicht und unter Berücksichtigung systematischer Fehlerquellen zu interpretieren. Falls Selbsteinschätzungen zur Erfassung der sprachlichen Kompetenzen im Deutschen eingesetzt werden, könnten Can-Do-Statements, die den Fähigkeitsbereich der Teilnehmenden angemessen abdecken, etwas besser geeignet sein als allgemein formulierte Selbsteinschätzungen, bei denen Teilnehmende ihre Kompetenzen auf einer Likert-Skala einstufen, insbesondere dann, wenn auch die Veränderung der Kompetenz über die Zeit erfasst werden soll. Items, die deutsche

Muttersprachlerinnen und Muttersprachler als Referenzgruppe vorgeben, haben die Genauigkeit der Selbsteinschätzungen der sprachlichen Kompetenzen im Deutschen der jugendlichen Flüchtlinge nicht verbessert gegenüber Items, die keine solche Referenz vorgeben. Dies ist wahrscheinlich auf die große Differenz zwischen den Deutschkompetenzen der beiden Gruppen zurückzuführen. Sofern eine passende Vergleichsgruppe gewählt werden kann, könnte dies die Genauigkeit der Selbsteinschätzungen erhöhen. Zehnstufige Endpunkt-gelabelte Antwortskalen erhöhten die Genauigkeit der Selbsteinschätzungen in dieser Stichprobe nicht gegenüber einer fünfstufigen vollständig gelabelten Antwortskala. Möglicherweise lässt sich die Genauigkeit der Selbsteinschätzungen durch die Verbesserung der Labels der fünfstufigen Skala jedoch erhöhen. Die vier untersuchten Arten von Selbsteinschätzungsitems unterschieden sich nicht grundlegend in ihrer Genauigkeit, weshalb für die Auswahl von Selbsteinschätzungsitems empfohlen wird, je nach untersuchter Stichprobe und Kontext ein geeignetes Maß zu wählen und die vorgeschlagenen Verbesserungen der Items umzusetzen und hinsichtlich ihrer Genauigkeit zu überprüfen.

Literaturverzeichnis

Abramson, L. Y., Metalsky, G. I. & Alloy, L. B. (1989). Hopelessness depression: A theory-based subtype of depression. *Psychological Review, 96*(2), 358–372. https://doi.org/10.1037/0033-295X.96.2.358

Ahadi, S. & Diener, E. (1989). Multiple determinants and effect size. *Journal of Personality and Social Psychology, 56*(3), 398–406. https://doi.org/10.1037/0022-3514.56.3.398

Albert, S. (1977). Temporal comparison theory. *Psychological Review, 84*(6), 485–503. https://doi.org/10.1037/0033-295X.84.6.485

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology, 49*(6), 1621–1630. https://doi.org/10.1037/0022-3514.49.6.1621

Alicke, M. D. & Govorun, O. (2005). The better-than-average effect. In M. D. Alicke, D. A. Dunning & J. I. Krueger (Hrsg.), *The self in social judgment* (S. 85–106). Taylor & Francis Group.

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J. & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology, 68*(5), 804. https://doi.org/10.1037/0022-3514.68.5.804

Alicke, M. D. & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology, 20*(1), 1–48. https://doi.org/10.1080/10463280802613866

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

Andersen, S. M. (1984). Self-knowledge and social inference: II. The diagnosticity of cognitive/affective and behavioral data. *Journal of Personality and Social Psychology, 46*(2), 294–307. https://doi.org/10.1037/0022-3514.46.2.294

Andersen, S. M. & Ross, L. (1984). Self-knowledge and social inference: I. The impact of cognitive/affective and behavioral data. *Journal of Personality and Social Psychology, 46*(2), 280–293. https://doi.org/10.1037/0022-3514.46.2.280

Appel, R. (1996). The lexicon in second language acquisition. In P. Jordens & J. Lalleman (Hrsg.), *Studies on language aquisition: Bd. 12. Investigating second language acquisition* (S. 381–403). Mouton de Gruyter.

Artelt, C. (2016). Teacher judgments and their role in the educational process. In R. Scott, M. Buchmann & S. Kosslyn (Hrsg.), *Emerging Trends in the Social and Behavioral Sciences* (S. 1–16). John Wiley & Sons.

Artelt, C. & Rausch, T. (2014). Accuracy of teacher judgments: When and for what reasons? In S. Krolak-Schwerdt, S. Glock & M. Böhmer (Hrsg.), *Teachers' professional development: Assessment, training, and learning* (S. 27–43). Sense Publishers.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Baumeister, R. F. (1989). The optimal margin of illusion. *Journal of Social and Clinical Psychology, 8*(2), 176–189. https://doi.org/10.1521/jscp.1989.8.2.176

Becker, R., Will, G. & Siegers, R. (2021). *Geflüchtete Jugendliche in der Sekundarstufe I – Ergebnisse der Befragung der institutionellen Kontextpersonen der ReGES-Studie* (LIfBi Working Paper No. 103). Bamberg. Leibniz-Institut für Bildungsverläufe. https://doi.org/10.5157/LIfBi:WP103:1.0

Beer, J. S., Chester, D. S. & Hughes, B. L. (2013). Social threat and cognitive load magnify self-enhancement and attenuate self-deprecation. *Journal of Experimental Social Psychology, 49*(4), 706–711. https://doi.org/10.1016/j.jesp.2013.02.017

Bem, D. J. (1972). Self-Perception Theory. In L. Berkowitz (Hrsg.), *Advances in experimental social psychology* (Bd. 6, S. 1–62). Academic Press. https://doi.org/10.1016/S0065-2601(08)60024-6

Ben-Shachar, M. S., Lüdecke, D. & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software, 5*(56), 2815. https://doi.org/10.21105/joss.02815

Berglas, S. & Jones, E. E. (1978). Drug choice as a self-handicapping strategy in response to noncontingent success. *Journal of Personality and Social Psychology, 36*(4), 405–417. https://doi.org/10.1037/0022-3514.36.4.405

Berry, J. W. (1997). Immigration, Acculturation, and Adaptation. *Applied Psychology: An International Review, 46*(1), 5–34. https://doi.org/10.1111/j.1464-0597.1997.tb01087.x

Biernat, M. (2003). Toward a broader view of social stereotyping. *American Psychologist, 58*(12), 1019–1027. https://doi.org/10.1037/0003-066X.58.12.1019

Biernat, M. (2005). *Standards and expectancies: Contrast and assimilation in judgments of self and others*. Taylor & Francis Group.

Biernat, M. & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology, 66*(1), 5–20. https://doi.org/10.1037/0022-3514.66.1.5

Biernat, M., Manis, M. & Nelson, T. E. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology, 60*(4), 485–499. https://doi.org/10.1037/0022-3514.60.4.485

Bishop, D. V. M. (1989). *Test for Reception of Grammar (TROG)*. Medical Research Council.

Blossfeld, H.-P. & Roßbach, H.-G. (Hrsg.). (2019). *Edition ZfE: Bd. 3. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (2. Aufl.). Springer VS. https://doi.org/10.1007/978-3-658-23162-0

Blumberg, H. H. (1972). Communication of interpersonal evaluations. *Journal of Personality and Social Psychology, 23*(2), 157–162. https://doi.org/10.1037/h0033027

Bollich, K. L., Johannet, P. M. & Vazire, S. (2011). In search of our true selves: Feedback as a path to self-knowledge. *Frontiers in Psychology, 2*(312), 1–6. https://doi.org/10.3389/fpsyg.2011.00312

Brantmeier, C. (2006). Advanced L2 learners and reading placement: Self-assessment, CBT, and subsequent performance. *System, 34*(1), 15–35. https://doi.org/10.1016/j.system.2005.08.004

Brantmeier, C., Vanderplank, R. & Strube, M. (2012). What about me? Individual self-assessment by skill and level of language instruction. *System, 40,* 144–160. https://doi.org/10.1016/j.system.2012.01.003

Brown, J. D. (1991). Accuracy and bias in self-knowledge. In C. R. Snyder (Hrsg.), *Handbook of social and clinical psychology: The health perspective* (S. 158–178). Pergamon Press.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte und erweiterte Aufl.). Pearson Studium.

Bundesamt für Migration und Flüchtlinge. (2020). *Migrationsbericht der Bundesregierung: Migrationsbericht 2018*. Nürnberg. Bundesamt für Migration und Flüchtlinge. https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/migration/migrationsbericht-2018.pdf?__blob=publicationFile&v=3

Camilleri, C. & Malewska-Peyre, H. (1997). Socialization and identity strategies. In J. W. Berry, P. R. Dasen & T. S. Saraswathi (Hrsg.), *Handbook of cross-cultural psychology (Vol. 2): Basic processes and human development* (2. Aufl., S. 41–67). Allyn and Bacon.

Cassidy, J., Ziv, Y., Mehta, T. G. & Feeney, B. C. (2003). Feedback seeking in children and adolescents: Associations with self-perceptions, attachment representations, and depression. *Child Development, 74*(2), 612–628. http://www.jstor.org/stable/3696334

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chambers, J. R. & Windschitl, P. D. (2004). Biases in social comparative judgments: the role of nonmotivated factors in above-average and comparative-optimism effects. *Psychological Bulletin, 130*(5), 813–838. https://doi.org/10.1037/0033-2909.130.5.813

Chen, C., Lee, S. & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, *6*(3), 170–175. https://doi.org/10.1111/j.1467-9280.1995.tb00327.x

Chiswick, B. R. & Miller, P. W. (2001). A model of destination-language acquisition: Application to male immigrants in Canada. *Demography*, *38*(3), 391–409. https://doi.org/10.1353/dem.2001.0025

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Colman, D. E. (2021). Characteristics of the judge that are related to accuracy. In T. D. Letzring & J. S. Spain (Hrsg.), *The Oxford Handbook of Accurate Personality Judgment* (S. 85–99). Oxford University Press.

Cooley, C. H. (1902). *Human nature and the social order*. Charles Scribner's Sons.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press. https://rm.coe.int/1680459f97

Creed, A. T. & Funder, D. C. (1998). The two faces of private self-consciousness: Self report, peer-report, and behavioral correlates. *European Journal of Personality*, *12*(6), 411–431. https://doi.org/10.1002/(SICI)1099-0984(199811/12)12:6<411::AID-PER317>3.0.CO;2-C

Cross, S. E. & Gore, J. S. (2012). Cultural models of the self. In M. R. Leary & J. P. Tangney (Hrsg.), *Handbook of self and identity* (2. Aufl., S. 587–614). The Guilford Press.

Cummins, R. A. & Gullone, E. (2000). *Why we should not use 5-point Likert scales: The case for subjective quality of life measurement* (Proceedings, Second International Conference on Quality of Life in Cities). Singapore. National University of Singapore.

Deutscher Volkshochschul-Verband e.V. (2011). *Programm-Management Sprachen: Ein Praxishandbuch zur Qualitätssicherung an Volkshochschulen*. Deutscher Volkshochschul-Verband e.V.

Dickhäuser, O. (2006). Fähigkeitsselbstkonzepte: Entstehung, Auswirkung, Förderung. *Zeitschrift für Pädagogische Psychologie*, *20*(1/2), 5–8. https://doi.org/10.1024/1010-0652.20.12.5

Dickhäuser, O. & Plenter, I. (2005). "Letztes Halbjahr stand ich zwei": Zur Akkuratheit selbst berichteter Noten. *Zeitschrift fur Padagogische Psychologie*, *19*(4), 219–224. https://doi.org/10.1024/1010-0652.19.4.219

Dufner, M., Gebauer, J. E., Sedikides, C. & Denissen, J. J. (2019). Self-enhancement and psychological adjustment: A meta-analytic review. *Personality and Social Psychology Review*, *23*(1), 48–72. https://doi.org/10.1177/1088868318756467

Dunn, L. M. & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test, Fourth edition*. Pearson.

Dunning, D. (2005). *Self-insight: Roadblocks and detours on the path to knowing thyself*. Taylor & Francis Group.

Dunning, D. & Helzer, E. G. (2014). Beyond the correlation coefficient in studies of self-assessment accuracy: Commentary on Zell & Krizan (2014). *Perspectives on Psychological Science, 9*(2), 126–130. https://doi.org/10.1177/1745691614521244

Dunning, D., Johnson, K., Ehrlinger, J. & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*(3), 83–87. https://doi.org/10.1111/1467-8721.01235

Dunning, D., Meyerowitz, J. A. & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57*(6), 1082–1090. https://doi.org/10.1037/0022-3514.57.6.1082

Edele, A., Kristen, C., Stanat, P. & Will, G. (2021). The education of recently arrived refugees in Germany: Conditions, processes, and outcomes. *Journal for Educational Research Online, 2021*(1), 5–15. https://doi.org/10.31244/jero.2021.01.01

Edele, A., Seuring, J., Kristen, C. & Stanat, P. (2015). Why bother with testing? The validity of immigrants' self-assessed language proficiency. *Social Science Research, 52*, 99–123. https://doi.org/10.1016/j.ssresearch.2014.12.017

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D. & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes, 105*(1), 98–121. https://doi.org/10.1016/j.obhdp.2007.05.002

Eid, M., Gollwitzer, M. & Schmitt, M. (2015). *Statistik und Forschungsmethoden* (4., überarbeitete und erweiterte Aufl.). Beltz.

Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion. Bachelorstudium Psychologie*. Hogrefe.

Epley, N. & Dunning, D. (2006). The mixed blessings of self-knowledge in behavioral prediction: Enhanced discrimination but exacerbated bias. *Personality and Social Psychology Bulletin, 32*(5), 641–655. https://doi.org/10.1177/0146167205284007

Esser, H. (2006). *Sprache und Integration: Die sozialen Bedingungen und Folgen des Spracherwerbs von Migranten*. Campus Verlag.

Falchikov, N. & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 395–430. https://doi.org/10.3102/00346543059004395

Felson, R. B. (1980). Communication barriers and the reflected appraisal process. *Social Psychology Quarterly, 43*(2), 223–233. https://doi.org/10.2307/3033625

Festinger, L. (1954). A theory of social comparison processes. *Human relations, 7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Filipp, S.-H. (1979). Entwurf eines heuristischen Bezugsrahmens für Selbstkonzept-Forschung: Menschliche Informationsverarbeitung und naive Handlungstheorie. In S.-H. Filipp (Hrsg.), *Selbstkonzept-Forschung: Probleme, Befunde, Perspektiven* (1. Aufl., S. 129–170). Klett-Cotta.

Filipp, S.-H. (2006). Kommentar zum Themenschwerpunkt: Entwicklung von Fähigkeitsselbstkonzepten. *Zeitschrift für Pädagogische Psychologie, 20*(1/2), 65–72. https://doi.org/10.1024/1010-0652.20.1.65

Filipp, S.-H. & Mayer, A.-K. (2005). Selbst und Selbst-Konzept. In H. Weber & T. Rammsayer (Hrsg.), *Handbuch der Persönlichkeitspsychologie und differentiellen Psychologie* (S. 266–276). Hogrefe Verlag.

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

Finnie, R. & Meng, R. (2005). Literacy and labour market outcomes: Self-assessment versus test score measures. *Applied Economics, 37*(17), 1935–1951. https://doi.org/10.1080/00036840500244519

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87*(6), 477–531. https://doi.org/10.1037/0033-295X.87.6.477

Fischer, K. W. & Bidell, T. R. (2006). Dynamic development of action and thought. In W. Damon & R. M. Lerner (Hrsg.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (6. Aufl., S. 313–399). Wiley. https://doi.org/10.1002/9780470147658.chpsy0107

Fischer, K. W., Hand, H. H., Watson, M. W., van Parys, M. M. & Tucker, J. L. (1984). Putting the child into socialization: The development of social categories in preschool children. In L. G. Katz (Hrsg.), *Current topics in early childhood education* (Bd. 5, S. 27–72). Ablex.

Fischer, L. & Durda, T. (2020). *NEPS Technical Report for receptive vocabulary: Scaling results of Starting Cohort 2 for kindergarten (wave 1), grade 1 (wave 3) and grade 3 (wave 5)* (NEPS Survey Paper No. 65). Bamberg. Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP65:1.0

Flink, C. & Park, B. (1991). Increasing consensus in trait judgments through outcome dependency. *Journal of Experimental Social Psychology, 27*(5), 453–467. https://doi.org/10.1016/0022-1031(91)90003-O

Fox, J. & Weisberg, S. (2019). *An R companion to applied regression* (3. Aufl.). SAGE. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Fox-Boyer, A. V. (2016). *TROG-D: Test zur Überprüfung des Grammatikverständnisses* (7. Auflage). Schulz-Kirchner Verlag.

Freund, P. A. & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin, 138*(2), 296–321. https://doi.org/10.1037/a0026556

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101*(1), 75–90. https://doi.org/10.1037/0033-2909.101.1.75

Funder, D. C. (1993). Judgments as data for personality and developmental psychology: Error versus accuracy. In D. C. Funder, R. D. Parke, C. Tomlinson-Keasey & K. Widaman (Hrsg.), *APA science volumes. Studying lives through time: Personality and development* (S. 121–146). American Psychological Association. https://doi.org/10.1037/10127-022

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. (2010). *The personality puzzle* (5. Aufl.). W. W. Norton & Company.

Funder, D. C. & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology, 55*(1), 149–158. https://doi.org/10.1037/0022-3514.55.1.149

Funder, D. C. & Colvin, C. R. (1997). Congruence of others' and self-judgments of personality. In R. Hogan, J. Johnson & S. R. Briggs (Hrsg.), *Handbook of personality psychology* (S. 617–647). Academic Press.

Gebauer, J. E., Wagner, J., Sedikides, C. & Neberich, W. (2013). Agency-communion and self-esteem relations are moderated by culture, religiosity, age, and sex: Evidence for the 'self-centrality breeds self-enhancement' principle. *Journal of Personality, 81*(3), 261–275. https://doi.org/10.1111/j.1467-6494.2012.00807.x

Gecas, V. (1982). The self-concept. *Annual Review of Sociology, 8*, 1–33. http://www.jstor.org/stable/2945986

Geiser, C. (2011). *Datenanalyse mit Mplus: Eine anwendungsorientierte Einführung* (2., durchgesehene Aufl.). VS Verlag für Sozialwissenschaften.

Gerber, J. P., Wheeler, L. & Suls, J. (2018). A social comparison theory meta-analysis 60+ years on. *Psychological Bulletin, 144*(2), 177–197. https://doi.org/10.1037/bul0000127

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology, 2*(1), 83–115. https://doi.org/10.1080/14792779143000033

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

Gnambs, T. (2017). *NEPS technical report for English reading competence: Scaling results of Starting Cohort 4 for grade 10* (NEPS Survey Paper No. 26). Bamberg. Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP26:1.0

Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I. & Cera, C. M. (2012). Self-ratings of spoken language dominance: A Multilingual Naming Test (MINT) and preliminary norms for young and aging Spanish–English bilinguals. *Bilingualism: Language and Cognition, 15*(03), 594–615. https://doi.org/10.1017/S1366728911000332

Green, J. D., Pinter, B. & Sedikides, C. (2005). Mnemic neglect and self-threat: Trait modifiability moderates self-protection. *European Journal of Social Psychology, 35*(2), 225–235. https://doi.org/10.1002/ejsp.242

Green, J. D., Sedikides, C. & Gregg, A. P. (2008). Forgotten but not gone: The recall and recognition of self-threatening memories. *Journal of Experimental Social Psychology, 44*(3), 547–561. https://doi.org/10.1016/j.jesp.2007.10.006

Gregg, A. P., Sedikides, C. & Gebauer, J. E. (2011). Dynamics of identity: Between self-enhancement and self-assessment. In S. J. Schwartz, K. Luyckx & V. L. Vignoles (Hrsg.), *Handbook of identity theory and research* (S. 305–327). Springer. https://doi.org/10.1007/978-1-4419-7988-9_14

Greve, W. (2000). Die Psychologie des Selbst - Konturen eines Forschungsthemas. In W. Greve (Hrsg.), *Psychologie des Selbst* (S. 15–36). Beltz, Psychologie Verlags Union.

Guo, J., Parker, P. D., Marsh, H. W. & Morin, A. J. S. (2015). Achievement, motivation, and educational choices: A longitudinal study of expectancy and value using a multiplicative perspective. *Developmental Psychology, 51*(8), 1163–1176. https://doi.org/10.1037/a0039440

Harter, S. (1990). Self and identity development. In S. S. Feldman & G. R. Elliott (Hrsg.), *At the threshold: The developing adolescent* (S. 352–387). Harvard Univ. Press.

Harter, S. (2006). The self. In N. Eisenberg, W. Damon & R. M. Lerner (Hrsg.), *Handbook of child psychology, Vol. 3: Social, emotional, and personality development* (6. Aufl., S. 505–570). John Wiley & Sons.

Harter, S. (2012). *The construction of the self: Developmental and sociocultural foundations* (2. Aufl.). The Guilford Press.

Harter, S. & Monsour, A. (1992). Developmental analysis of conflict caused by opposing attributes in the adolescent self-portrait. *Developmental Psychology, 28*(2), 251–260. https://doi.org/10.1037/0012-1649.28.2.251

Heine, S. J. & Hamamura, T. (2007). In search of East Asian self-enhancement. *Personality and Social Psychology Review*, *11*(1), 4–27. https://doi.org/10.1177/1088868306294587

Heine, S. J., Kitayama, S. & Hamamura, T. (2007). Inclusion of additional studies yields different conclusions: Comment on Sedikides, Gaertner, & Vevea (2005), Journal of Personality and Social Psychology. *Asian Journal of Social Psychology*, *10*(2), 49–58. https://doi.org/10.1111/j.1467-839X.2007.00211.x

Heine, S. J., Lehman, D. R., Peng, K. & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group effect. *Journal of Personality and Social Psychology*, *82*(6), 903–918. https://doi.org/10.1037/0022-3514.82.6.903

Helm, F., Marsh, H. W., Dicke, T. & Möller, J. (2020). Dimensional comparison theory. In J. Suls, R. L. Collins & L. Wheeler (Hrsg.), *Social Comparison, Judgment, and Behavior* (S. 201–225). Oxford University Press. https://doi.org/10.1093/oso/9780190629113.003.0008

Herreen, D. & Zajac, I. T. (2018). The reliability and validity of a self-report measure of cognitive abilities in older adults: More personality than cognitive function. *Journal of Intelligence*, *6*(1), 1–15. https://doi.org/10.3390/jintelligence6010001

Higgins (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review*, *94*(3), 319–340. https://doi.org/10.1037/0033-295X.94.3.319

Ho, D. Y. (1976). On the concept of face. *American Journal of Sociology*, *81*(4), 867–884. http://www.jstor.org/stable/2777600

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Sage Publications.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2. Aufl.). Sage Puiblications.

Huguet, P., Dumas, F., Marsh, H. W., Régner, I., Wheeler, L., Suls, J., Seaton, M. & Nezlek, J. (2009). Clarifying the role of social comparison in the big-fish-little-pond effect (BFLPE): An integrative study. *Journal of Personality and Social Psychology*, *97*(1), 156–170. https://doi.org/10.1037/a0015558

James, W. (1890). *The principles of psychology* (I). Henry Holt.

John, O. P. & Robins, R. W. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology*, *66*(1), 206–219. https://doi.org/10.1037/0022-3514.66.1.206

Karcher, M. J. & Fischer, K. W. (2004). A developmental sequence of skills in adolescents' intergroup understanding. *Journal of Applied Developmental Psychology*, *25*(3), 259–282. https://doi.org/10.1016/j.appdev.2004.04.001

Kievit, R. A., Brandmaier, A. M., Ziegler, G., van Harmelen, A.-L., Mooij, S. M. M. de, Moutoussis, M., Goodyer, I. M., Bullmore, E., Jones, P. B., Fonagy, P., the NSPN

Consortium, Lindenberger, U. & Dolan, R. J. (2018). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *Developmental Cognitive Neuroscience, 33*, 99–117. https://doi.org/10.1016/j.dcn.2017.11.007

Klar, Y. (2002). Way beyond compare: Nonselective superiority and inferiority biases in judging randomly assigned group members relative to their peers. *Journal of Experimental Social Psychology, 38*(4), 331–351. https://doi.org/10.1016/S0022-1031(02)00003-3

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4. Aufl.). The Guilford Press.

Koppel, L., Andersson, D., Tinghög, G., Västfjäll, D. & Feldman, G. (2021). We are all less risky and more skillful than our fellow drivers: Replication and extension of Svenson (1981). Vorab-Onlinepublikation. https://doi.org/10.17605/OSF.IO/FXPWB

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Krosnick, J. A. & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Hrsg.), *Handbook of survey research* (2. Aufl., S. 263–313). Emerald Group Publishing Limited.

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology, 77*(2), 221–232. https://doi.org/10.1037/0022-3514.77.2.221

Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Krzyzaniak, S. L. & Letzring, T. D. (2021). Characteristics of traits that are related to accuracy of personality judgments. In T. D. Letzring & J. S. Spain (Hrsg.), *The Oxford Handbook of Accurate Personality Judgment* (S. 119–131). Oxford University Press.

Kwang, T. & Swann, W. B. (2010). Do people embrace praise even when they feel unworthy? A review of critical tests of self-enhancement versus self-verification. *Personality and Social Psychology Review, 14*(3), 263–280. https://doi.org/10.1177/1088868310365876

Lang, F. R., Kamin, S., Rohr, M., Stünkel, C. & Willinger, B. (2014). *Erfassung der fluiden kognitiven Leistungsfähigkeit über die Lebensspanne im Rahmen der National Educational Panel Study: Abschlussbericht zu einer NEPS-Ergänzungsstudie* (NEPS Working Paper No. 43). Bamberg. Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel. https://www.lifbi.de/Portals/0/Working%20Papers/WP_XLIII.pdf

Leary, M. R. (2007). Motivational and emotional aspects of the self. *Annual Review of Psychology, 58*(1), 317–344. https://doi.org/10.1146/annurev.psych.58.110405.085658

Leary, M. R. & Tangney, J. P. (2012). The self as an organizing construct in the behavioral and social sciences. In M. R. Leary & J. P. Tangney (Hrsg.), *Handbook of self and identity* (2. Aufl., S. 1–18). The Guilford Press.

LeBlanc, R. & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly, 19*(4), 673–687. https://doi.org/10.2307/3586670

Lecky, P. (1945/1969). *Self-consistency: A theory of personality* (J. F. A. Taylor & F. C. Thorne, Hg.). Anchor Books. (Erstveröffentlichung 1945)

Leibniz-Institut für Bildungsverläufe e.V. (2018). *Erhebungsinstrumente (SUF-Version): NEPS Startkohorte 2 - Kindergarten: Frühe Bildung in Kindergarten und Grundschule: Welle 4 - 4.0.0*. Bamberg. https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC2/4-0-0/SC2_4-0-0_W4_de.pdf

Lenhard, A., Lenhard, W., Segerer, R. & Suggate, S. (2015). *Peabody Picture Vocabulary Test - 4. Ausgabe: Deutsche Fassung*. Pearson.

Letzring, T. D. & Funder, D. C. (2021). The Realistic Accuracy Model. In T. D. Letzring & J. S. Spain (Hrsg.), *The Oxford Handbook of Accurate Personality Judgment* (S. 9–22). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190912529.013.2

Leung, S. O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of Social Service Research, 37*(4), 412–421. https://doi.org/10.1080/01488376.2011.580697

Leung, S. O. & Wong, P. M. (2008). Validity and reliability of Chinese Rosenberg Self-Esteem Scale. *New Horizons in Education, 56*(1), 62–69.

Lin, L.-M., Moore, D. & Zabrucky, K. M. (2001). An assessment of students' calibration of comprehension and calibration of performance using multiple measures. *Reading Psychology, 22*(2), 111–128. https://doi.org/10.1080/02702710119125

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., Rioux, C., Odejimi, O. A. & Stickley, Z. L. (2022). *Parceling in Structural Equation Modeling: A comprehensive introduction for developmental scientists*. Cambridge University Press. https://doi.org/10.1017/9781009211659

Lorenz, C., Berendes, K. & Weinert, S. (2017). *Measuring receptive grammar in kindergarten and elementary school children in the German National Educational Panel Study* (NEPS Survey Paper No. 24). Bamberg. Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP24:1.0

Lu, R., Bao, N., Zhang, X. & Shi, J. (2022). Attentional resource allocation among individuals with different fluid intelligence: The integrated control hypothesis and its evidence from pupillometry. *Neuropsychologia, 169*, 108190. https://doi.org/10.1016/j.neuropsychologia.2022.108190

Mabe, P. A. & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*(3), 280–296. https://doi.org/10.1037/0021-9010.67.3.280

Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology, 35*(2), 63–78. https://doi.org/10.1037/0022-3514.35.2.63

Markus, H. & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*(2), 224–253. https://doi.org/10.1037/0033-295X.98.2.224

Markus, H. & Nurius, P. (1986). Possible selves. *American Psychologist, 41*(9), 954–969. https://doi.org/10.1037/0003-066X.41.9.954

Markus, H. & Wurf, E. (1987). The dynamic self-concept: A social psychological perspective. *Annual Review of Psychology, 38*(1), 299–337. https://doi.org/10.1146/annurev.ps.38.020187.001503

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23*(1), 129–149. https://doi.org/10.3102/00028312023001129

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79*(3), 280–295. https://doi.org/10.1037/0022-0663.79.3.280

Marsh, H. W. & Parker, J. W. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology, 47*(1), 213–231. https://doi.org/10.1037/0022-3514.47.1.213

Marsh, H. W. & Shavelson, R. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational psychologist, 1520*(20), 107–125. https://doi.org/10.1207/s15326985ep2003

Marsh, H. W. & Yeung, A. S. (1997). Causal effects of academic self-concept on academic achievement: Structural equation models of longitudinal data. *Journal of Educational Psychology, 89*(1), 41–54. https://doi.org/10.1037/0022-0663.89.1.41

Mather, M. (2006). Why memories may become more positive as people age. In B. Uttl, N. Ohta & A. L. Siegenthaler (Hrsg.), *Memory and emotion: Interdisciplinary perspectives* (S. 135–158). Blackwell Publishing. https://doi.org/10.1002/9780470756232.ch7

McCrae, R. R. (1982). Consensual validation of personality traits: Evidence from self-reports and ratings. *Journal of Personality and Social Psychology, 43*(2), 293–303. https://doi.org/10.1037/0022-3514.43.2.293

McCrae, R. R. & Costa, P. T. (1995). Trait explanations in personality psychology. *European Journal of Personality, 9*(4), 231–252. https://doi.org/10.1002/per.2410090402

Merkin, R. & Ramadan, R. (2010). Facework in Syria and the United States: A cross-cultural comparison. *International Journal of Intercultural Relations, 34*(6), 661–669. https://doi.org/10.1016/j.ijintrel.2010.05.006

Mezulis, A. H., Abramson, L. Y., Hyde, J. S. & Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin, 130*(5), 711–747. https://doi.org/10.1037/0033-2909.130.5.711

Mignault, M.-C. & Human, L. J. (2021). The good target of personality judgments. In T. D. Letzring & J. S. Spain (Hrsg.), *The Oxford Handbook of Accurate Personality Judgment* (S. 100–118). Oxford University Press.

Min, I., Cortina, K. S. & Miller, K. F. (2016). Modesty bias and the attitude-achievement paradox across nations: A reanalysis of TIMSS. *Learning and Individual Differences, 51*, 359–366. https://doi.org/10.1016/j.lindif.2016.09.008

Möller, J., Helm, F., Müller-Kalthoff, H., Nagy, N. & Marsh, H. W. (2015). Dimensional comparisons and their consequences for self-concept, motivation, and emotion. In J. D. Wright (Hrsg.), *International encyclopedia of the social & behavioral sciences* (2. Aufl., S. 430–436). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.26092-3

Möller, J. & Köller, O. (2001). Frame of reference effects following the announcement of exam results. *Contemporary Educational Psychology, 26*, 277–287. https://doi.org/10.1006/ceps.2000.1055

Möller, J. & Marsh, H. W. (2013). Dimensional comparison theory. *Psychological Review, 120*(3), 544–560. https://doi.org/10.1037/a0032459

Möller, J., Pohlmann, B., Köller, O. & Marsh, H. W. (2009). A meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79*(3), 1129–1167. https://doi.org/10.3102/0034654309337522

Möller, J., Zitzmann, S., Helm, F., Machts, N. & Wolff, F. (2020). A meta-analysis of relations between achievement and self-concept. *Review of Educational Research, 90*(3), 376–419. https://doi.org/10.3102/0034654320919354

Mummendey, H. D. (2006). *Psychologie des 'Selbst': Theorien, Methoden und Ergebnisse der Selbstkonzeptforschung*. Hogrefe.

Neubauer, A. C. & Hofer, G. (2019). Self- and other-estimates of intelligence. In R. J. Sternberg (Hrsg.), *The Cambridge Handbook of Intelligence* (2. Aufl., S. 1179–1200). Cambridge University Press. https://doi.org/10.1017/9781108770422

Niehues, W., Rother, N. & Siegert, M. (04/2021). *Vierte Welle der IAB-BAMF-SOEP-Befragung von Geflüchteten: Spracherwerb und soziale Kontakte schreiten bei Geflüchteten voran* (BAMF-Kurzanalysen). Nürnberg. Forschungszentrum Migration, Integration und Asyl des Bundesamtes für Migration und Flüchtlinge. https://www.bamf.de/SharedDocs/Anlagen/DE/Forschung/Kurzanalysen/kurzanalyse4-2021_iab-bamf-soep-befragung-4te-welle.pdf

Niepel, C., Marsh, H. W., Guo, J., Pekrun, R. & Möller, J. (2022). Revealing dynamic relations between mathematics self-concept and perceived achievement from lesson to lesson: An experience-sampling study. *Journal of Educational Psychology, 114*(6), 1380–1393. https://doi.org/10.1037/edu0000716

Obry, M., Schild, A., Will, G. & Kopp, F. (2021). *Die Messung des rezeptiven Wortschatzes in der Flüchtlingsstudie ReGES (Welle 1)* (LIfBi Working Paper No. 98). Leibniz-Institut für Bildungsverläufe. https://doi.org/10.5157/LIfBi:WP98:2.0

Pahl, S. & Eiser, J. R. (2007). How malleable is comparative self-positivity? The effects of manipulating judgemental focus and accessibility. *European Journal of Social Psychology, 37*(4), 617–627. https://doi.org/10.1002/ejsp.372

Pohl, S. & Carstensen, C. H. (2012). *NEPS Technical Report - Scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg. Otto-Friedrich-Universität, Nationales Bildungspanel. https://www.lifbi.de/Portals/0/Working%20Papers/WP_XIV.pdf

Pohlmann, B., Möller, J. & Streblow, L. (2006). Zur Bedeutung dimensionaler Aufwärts- und Abwärtsvergleiche. *Zeitschrift für Pädagogische Psychologie, 20*(1/2), 19–25. https://doi.org/10.1024/1010-0652.20.12.19

R Core Team. (2020). *R: A language and environment for statistical computing*. https://www.r-project.org/

Revelle, W. (2022). *psych: Procedures for Personality and Psychological Research* (Version 2.2.5) [Computer software]. Northwestern University. Evanston, Illinois. https://CRAN.R-project.org/package=psych

Rosenberg, M. (1979). *Conceiving the self*. Basic Books.

Ross, L. & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. Temple University Press.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*(1), 1–20. https://doi.org/10.1177/026553229801500101

Ruble, D. N. & Frey, K. S. (1991). Changing patterns of comparative behavior as skills are aquired: A functional model of self-evaluation. In J. Suls & T. A. Wills (Hrsg.), *Social comparison: Contemporary theory and research* (S. 79–113). L. Erlbaum Associates.

Ruland, M., Sandbrink, K., Cohrs, I. & Hess, D. (2020). *Methodenbericht: ReGES - Befragung C10*. Bonn. infas Institut für angewandte Sozialwissenschaft. https://www.reges-data.de/Portals/4/Datenzentrum/Dokumentation/General/ReGES_Methodenbericht_W7_C10.pdf

Ruland, M., Steinwede, A., Sandbrink, K., Lesaar, S. & Hess, D. (2019). *Methodenbericht: ReGES-Erstbefragung C04*. Bonn. infas Institut für angewandte Sozialwissenschaft GmbH. https://www.reges-data.de/Portals/4/Datenzentrum/Dokumentation/General/ReGES_Methodenbericht_W1_C04.pdf

Schkade, D. A. & Kahneman, D. (1998). Does living in California make people happy? A focusing illusion in judgments of life satisfaction. *Psychological Science*, *9*(5), 340–346. https://doi.org/10.1111/1467-9280.00066

Schneider, W. J. & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Hrsg.), *Contemporary intellectual assessment: Theories, tests, and issues* (3. Aufl., S. 99–144). Guilford Press.

Schneider, W. (1998). Performance prediction in young children: Effects of skill, metacognition and wishful thinking. *Developmental Science*, *1*(2), 291–297. https://doi.org/10.1111/1467-7687.00044

Schütz, A. & Baumeister, R. F. (2017). Positive illusions and the happy mind. In M. D. Robinson & M. Eid (Hrsg.), *The happy mind: Cognitive contributions to well-being* (S. 177–193). Springer International Publishing. https://doi.org/10.1007/978-3-319-58763-9_10

Schütz, A., Fehn, T. & Baumeister, R. F. (2018). Self. In V. Zeigler-Hill & T. K. Shackelford (Hrsg.), *Encyclopedia of personality and individual differences*. Springer International Publishing. https://doi.org/10.1007/978-3-319-28099-8_1998-1

Schütz, A., Rüdiger, M. & Rentzsch, K. (2016). *Lehrbuch Persönlichkeitspsychologie*. Hogrefe Verlag.

Schwinger, M., Wirthwein, L., Lemmer, G. & Steinmayr, R. (2014). Academic self-handicapping and achievement: A meta-analysis. *Journal of Educational Psychology*, *106*(3), 744–761. https://doi.org/10.1037/a0035832

Sedikides, C. (1993). Assessment, enhancement, and verification determinants of the self-evaluation process. *Journal of Personality and Social Psychology*, *65*(2), 317–338. https://doi.org/10.1037/0022-3514.65.2.317

Sedikides, C. (2009). On self-protection and self-enhancement regulation: The role of self-improvement and social norms. In J. P. Forgas, R. F. Baumeister & D. M. Tice (Hrsg.), *Psychology of self-regulation: cognitive, affective, and motivational processes* (S. 73–92). Psychology Press.

Sedikides, C. & Alicke, M. D. (2019). The five pillars of self-enhancement and self-protection. In R. M. Ryan (Hrsg.), *Oxford handbook of human motivation* (2. Aufl., S. 307–319). Oxford University Press.

Sedikides, C. & Green, J. D. (2004). What I don't recall can't hurt me: Information negativity versus information inconsistency as determinants of memorial self-defense. *Social Cognition, 22*(1), 4–29. https://doi.org/10.1521/soco.22.1.4.30987

Sedikides, C. & Green, J. D. (2009). Memory as a self-protective mechanism. *Social and Personality Psychology Compass, 3*(6), 1055–1068. https://doi.org/10.1111/j.1751-9004.2009.00220.x

Sedikides, C., Green, J. D., Saunders, J., Skowronski, J. J. & Zengel, B. (2016). Mnemic neglect: Selective amnesia of one's faults. *European Review of Social Psychology, 27*(1), 1–62. https://doi.org/10.1080/10463283.2016.1183913

Sedikides, C., Herbst, K. C., Hardin, D. P. & Dardis, G. J. (2002). Accountability as a deterrent to self-enhancement: The search for mechanisms. *Journal of Personality and Social Psychology, 83*(3), 592–605. https://doi.org/10.1037/0022-3514.83.3.592

Selman, R. L. (1980). *The growth of interpersonal understanding: Developmental and clinical analyses*. Academic Press.

Shavelson, R. J., Hubner, J. J. & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research, 46*(3), 407–441. https://doi.org/10.3102/00346543046003407

Shrauger, J. S. & Schoeneman, T. J. (1979). Symbolic interactionist view of self-concept: Through the looking glass darkly. *Psychological Bulletin, 86*(3), 549–573. https://doi.org/10.1037/0033-2909.86.3.549

Skaalvik, E. M. & Skaalvik, S. (2002). Internal and external frames of reference for academic self-concept. *Educational psychologist, 37*(4), 233–244. https://doi.org/10.1207/S15326985EP3704_3

Spinath, B. & Spinath, F. M. (2005). Development of self-perceived ability in elementary school: the role of parents' perceptions, teacher evaluations, and intelligence. *Cognitive Development, 20*(2), 190–204. https://doi.org/10.1016/j.cogdev.2005.01.001

Steinhauer, H. W., Zinn, S. & Will, G. (2019). Sampling refugees for an educational longitudinal survey. *Survey Methods: Insights from the Field*, 1–17. https://doi.org/10.13094/SMIF-2019-00007

Stevenson, H. W. & Zusho, A. (2002). Adolescence in China and Japan: Adapting to a changing environment. In B. Bradford Brown, R. W. Larson & T. S. Saraswathi (Hrsg.), *The world's youth: Adolescence in eight regions of the globe* (S. 141–170). Cambridge University Press. https://doi.org/10.1017/CBO9780511613814.006

Strack, F. & Martin, L. L. (1987). Thinking, judging, and communicating: A process account of context effects in attitude surveys. In H.-J. Hippler, N. Schwarz & S. Sudman (Hrsg.), *Social information processing and survey methodology* (S. 123–148). Springer-Verlag.

Südkamp, A., Kaiser, J. & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–762. https://doi.org/10.1037/a0027627

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica, 47*(2), 143–148. https://doi.org/10.1016/0001-6918(81)90005-6

Swann, W. B. (1983). Self-verification: Bringing social reality into harmony with the self. In J. Suls & A. G. Greenwald (Hrsg.), *Psychological perspectives on the self* (Bd. 2, S. 33–66). Erlbaum.

Swann, W. B. & Buhrmester, M. D. (2012). Self-verification: The search for coherence. In M. R. Leary & J. P. Tangney (Hrsg.), *Handbook of self and identity* (2. Aufl., S. 405–424). The Guilford Press.

Swann, W. B., Griffin, J. J., Predmore, S. C. & Gaines, B. (1987). The cognitive–affective crossfire: When self-consistency confronts self-enhancement. *Journal of Personality and Social Psychology, 52*(5), 881–889. https://doi.org/10.1037/0022-3514.52.5.881

Swann, W. B., Hixon, J. G., Stein-Seroussi, A. & Gilbert, D. T. (1990). The fleeting gleam of praise: Cognitive processes underlying behavioral reactions to self-relevant feedback. *Journal of Personality and Social Psychology, 59*(1), 17–26. https://doi.org/10.1037/0022-3514.59.1.17

Swann, W. B. & Read, S. J. (1981). Self-verification processes: How we sustain our self-conceptions. *Journal of Experimental Social Psychology, 17*(4), 351–372. https://doi.org/10.1016/0022-1031(81)90043-3

Taylor, S. E. & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*(2), 193–210.

Taylor, S. E., Neter, E. & Wayment, H. A. (1995). Self-evaluation processes. *Personality and Social Psychology Bulletin, 21*(12), 1278–1287. https://doi.org/10.1177/01461672952112005

TNS Infratest Sozialforschung. (2016). *Erhebungsinstrumente der IAB-BAMF-SOEP-Befragung von Geflüchteten 2016: Integrierter Personen- und Biografiefragebogen, Stichproben M3-M4* (SOEP Survey Papers 362: Series A). Berlin. DIW/SOEP. http://hdl.handle.net/10419/183118

Tourangeau, R., Rips, L. J. & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K. & Crowther, D. (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition, 19*(1), 122–140. https://doi.org/10.1017/S1366728914000832

Trope, Y. (1975). Seeking information about one's ability as a determinant of choice among tasks. *Journal of Personality and Social Psychology, 32*(6), 1004–1013. https://doi.org/10.1037//0022-3514.32.6.1004

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

Trzesniewski, K. H., Kinal, M. P.-A. & Donnellan, M. B. (2011). Self-enhancement and self-protection in a developmental context. In M. D. Alicke & C. Sedikides (Hrsg.), *Handbook of self-enhancement and self-protection* (S. 341–357). The Guilford Press.

van der Meer, E., Beyer, R., Horn, J., Foth, M., Bornemann, B., Ries, J., Kramer, J., Warmuth, E., Heekeren, H. R. & Wartenburger, I. (2010). Resource allocation and fluid intelligence: Insights from pupillometry. *Psychophysiology, 47*(1), 158–169. https://doi.org/10.1111/j.1469-8986.2009.00884.x

van der Westhuizen, L., Arens, A. K., Greiff, S., Fischbach, A. & Niepel, C. (2022). The generalized internal/external frame of reference model with academic self-concepts, interests, and anxieties in students from different language backgrounds. *Contemporary Educational Psychology, 68*, 102037. https://doi.org/10.1016/j.cedpsych.2021.102037

van Lange, P. A. M. & Sedikides, C. (1998). Being more honest but not necessarily more intelligent than others: generality and explanations for the Muhammad Ali effect. *European Journal of Social Psychology, 28*(4), 675–680. https://doi.org/10.1002/(SICI)1099-0992(199807/08)28:4<675::AID-EJSP883>3.0.CO;2-5

Wheeler, L., Martin, R. & Suls, J. (1997). The proxy model of social comparison for self-assessment of ability. *Personality and Social Psychology Review, 1*(1), 54–61. https://doi.org/10.1207/s15327957pspr0101_4

Will, G., Balaban, E., Dröscher, A., Homuth, C. & Welker, J. (2018). *Integration von Flüchtlingen in Deutschland: Erste Ergebnisse aus der ReGES-Studie* (Aktualisierung LIfBi Working Paper No. 76). Bamberg. Leibniz-Institut für Bildungsverläufe. https://www.lifbi.de/Portals/13/LIfBi%20Working%20Papers/Aktualisierung_WP_LXXVI.pdf

Will, G., Becker, R. & Weigand, D. (2020). COVID-19 lockdown during field work: Challenges and strategies in continuing the ReGES study. *Survey Research Methods, 14*(2), 247–252. https://doi.org/10.18148/srm/2020.v14i2.7753

Will, G., Homuth, C., Maurice, J. von & Roßbach, H. G. (2021). Integration of recently arrived underage refugees: Research potential of the Study ReGES - Refugees in the German Educational System. *European Sociological Review, 37*(6), 1027–1043. https://doi.org/10.1093/esr/jcab033

Zell, E. & Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science*, *9*(2), 111–125. https://doi.org/10.1177/1745691613518075

Zell, E., Strickhouser, J. E., Sedikides, C. & Alicke, M. D. (2020). The better-than-average effect in comparative self-evaluation: A comprehensive review and meta-analysis. *Psychological Bulletin*, *146*(2), 118–149. https://doi.org/10.1037/bul0000218

Anhang A – Itemkennwerte Tests und Skalen

Tabelle A 1. Schwierigkeiten und Trennschärfen der Items des PPVT-4 in der siebten Erhebungswelle

Setnr.	Itemnr.	Schwierigkeit	Trennschärfe	Setnr.	Itemnr.	Schwierigkeit	Trennschärfe
1	1	1.00	.09	9	1	.51	.39
1	2	.98	.21	9	2	.23	.38
1	3	1.00	.03	9	3	.58	.33
1	4	.94	.33	9	4	.50	.56
1	5	.97	.26	9	5	.29	.40
1	6	.98	.21	9	6	.54	.65
1	7	.99	.18	9	7	.50	.59
1	8	.93	.35	9	8	.72	.60
1	9	1.00	.11	9	9	.57	.49
1	10	.92	.36	9	10	.68	.50
1	11	.99	.17	9	11	.64	.47
1	12	.96	.32	9	12	.75	.65
2	1	.89	.42	10	1	.42	.51
2	2	.99	.14	10	2	.11	.31
2	3	.96	.32	10	3	.31	.55
2	4	.99	.14	10	4	.65	.68
2	5	.99	.13	10	5	.12	.35
2	6	.98	.20	10	6	.59	.61
2	7	.95	.27	10	7	.43	.60
2	8	1.00	.07	10	8	.56	.64
2	10	.99	.14	10	9	.64	.72
2	11	1.00	.02	10	10	.31	.36
2	12	.94	.31	10	11	.57	.66
3	1	.99	.15	10	12	.31	.47
3	2	1.00	.10	11	1	.55	.74

3	3	1.00	.13	11	2	.49	.67
3	4	.83	.40	11	3	.34	.50
3	5	.99	.16	11	4	.26	.54
3	6	1.00	.11	11	5	.48	.69
3	7	.96	.24	11	6	.23	.50
3	8	.97	.23	11	7	.15	.30
3	9	.99	.11	11	8	.24	.55
3	10	.76	.45	11	9	.18	.39
3	11	.97	.24	11	10	.22	.41
3	12	.99	.15	11	11	.49	.72
4	1	.54	.53	11	12	.31	.59
4	2	.98	.17	12	1	.24	.54
4	3	.79	.34	12	2	.06	.26
4	4	.97	.24	12	3	.03	.24
4	5	.87	.42	12	4	.46	.73
4	6	.89	.43	12	5	.50	.80
4	7	.81	.34	12	6	.10	.37
4	8	.97	.29	12	7	.36	.69
4	9	.64	.46	12	8	.20	.51
4	10	1.00	.10	12	9	.30	.60
4	11	.97	.23	12	10	.28	.51
4	12	.75	.42	12	11	.44	.77
5	1	.45	.35	12	12	.39	.69
5	2	.61	.28	13	1	.09	.38
5	3	.99	.11	13	2	.21	.59
5	4	.99	.20	13	3	.05	.27
5	5	.50	.55	13	4	.36	.73
5	6	.84	.51	13	5	.18	.54
5	7	.94	.29	13	6	.21	.47

5	8	.80	.43	13	7	.24	.63
5	9	.98	.24	13	8	.40	.75
5	10	.85	.38	13	9	.10	.35
5	11	.68	.51	13	10	.20	.51
5	12	.82	.54	13	11	.21	.52
6	1	.84	.51	13	12	.25	.57
6	2	.97	.33	14	1	.11	.42
6	3	.50	.38	14	2	.17	.59
6	4	.97	.33	14	3	.17	.61
6	5	.66	.42	14	4	.22	.63
6	6	.56	.49	14	5	.25	.70
6	7	.75	.52	14	6	.17	.50
6	8	.89	.39	14	7	.25	.67
6	9	.39	.39	14	8	.14	.49
6	10	.72	.53	14	9	.09	.39
6	11	.70	.48	14	10	.05	.25
6	12	.86	.45	14	11	.20	.63
7	1	.94	.46	14	12	.14	.54
7	2	.62	.54	15	1	.05	.38
7	3	.94	.45	15	2	.15	.60
7	4	.90	.45	15	3	.15	.58
7	5	.66	.42	15	4	.12	.51
7	6	.55	.38	15	5	.12	.53
7	7	.64	.48	15	6	.12	.49
7	8	.87	.51	15	7	.12	.54
7	9	.49	.61	15	8	.17	.59
7	10	.86	.45	15	9	.14	.55
7	11	.60	.34	15	10	.13	.51
7	12	.58	.27	15	11	.09	.43

8	1	.32	.16	15	12	.13	.52
8	2	.40	.39	16	1	.09	.48
8	3	.64	.57	16	2	.13	.57
8	4	.68	.55	16	3	.06	.38
8	5	.92	.46	16	4	.08	.45
8	6	.79	.49	16	5	.05	.36
8	7	.48	.38	16	6	.03	.23
8	8	.75	.40	16	7	.04	.32
8	9	.60	.55	16	8	.09	.46
8	10	.55	.47	16	9	.11	.52
8	11	.31	.61	16	10	.03	.24
8	12	.83	.50	16	11	.10	.49
				16	12	.04	.31

Anmerkungen. Die Schwierigkeit entspricht dem Anteil korrekter Antworten. Die Trennschärfe entspricht der Korrelation des Items mit dem um das Item korrigierten Summenwert. Schwierigkeit und Trennschärfe wurden nur für die ersten 16 Sets des PPVT-4 berechnet, da nur wenige Teilnehmende die Sets 17 bis 19 bearbeitet haben. Das neunte Item in Set 2 wurde von den Analysen zur Schwierigkeit und Trennschärfe ausgeschlossen, da es keine Varianz hatte. Nicht bearbeitete Items oberhalb des Deckensets wurden als falsch gelöst kodiert, nicht bearbeitete Items unterhalb des Bodensets wurden als richtig gelöst kodiert. $n = 773$.

Tabelle A 2. Schwierigkeiten und Trennschärfen der Items des TROG-D in beiden Erhebungswellen

		Erhebungswelle 1				Erhebungswelle 7			
Itemblock	**Itemnummer**	**Anzahl gültiger Werte**	**Schwierigkeit**	**Trennschärfe**	**Trennschärfe 47 Items**	**Anzahl gültiger Werte**	**Schwierigkeit**	**Trennschärfe**	**Trennschärfe 47 Items**
A	1	1716	.99	.14	.14	666	.99	.27	.27
A	2	1786	.98	.26	.26	673	.98	.23	.23
A	3	1785	.92	.29	.29	671	.96	.30	.30
A	4	1791	1.00	.09	.09	671	.99	.33	.33
B	1	1792	.99	.14	.14	674	.99	.34	.34
B	2	1780	.63	.39	.39	673	.76	.45	.45
B	3	1788	.98	.20	.21	671	.99	.24	.24

		Erhebungswelle 1				Erhebungswelle 7			
Item-block	Item-nummer	Anzahl gültiger Werte	Schwierig-keit	Trenn-schärfe	Trenn-schärfe 47 Items	Anzahl gültiger Werte	Schwierig-keit	Trenn-schärfe	Trenn-schärfe 47 Items
B	4	1789	.98	.18	.18	673	.98	.21	.21
C	1	1784	.95	.20	.21	669	.97	.30	.31
C	2	1796	.95	.36	.36	672	.98	.31	.32
C	3	1777	.94	.16	.16	667	.96	.25	.25
C	4	1795	.99	.23	.24	671	.99	.32	.33
D	1	1789	.98	.23	.23	673	.98	.31	.32
D	3	1790	.99	.22	.22	671	1.00	.27	.27
E	3	1783	.76	.48	.48	675	.88	.48	.48
E	4	1781	.80	.51	.51	673	.90	.46	.46
F	3	1791	.96	.24	.25	674	.98	.26	.27
F	4	1791	.96	.26	.26	673	.98	.24	.25
G	1	1786	.92	.34	.34	672	.95	.39	.40
G	4	1790	.93	.39	.39	674	.96	.42	.43
H	1	1789	.48	.23	.23	675	.52	.16	.16
H	3	1792	.89	.42	.42	673	.96	.38	.39
I	1	1789	.83	.49	.49	670	.91	.47	.48
I	3	1789	.90	.34	.34	674	.94	.21	.21
J	2	1774	.78	.35	.36	672	.83	.33	.34
J	3	1791	.80	.42	.42	672	.90	.44	.44
K	1	1775	.61	.44	.45	673	.77	.51	.52
K	4	1782	.50	.43	.43	671	.68	.48	.48
L	1	1782	.81	.26	.27	672	.87	.25	.25
L	2	1790	.72	.51	.51	672	.82	.51	.51
M	2	1781	.66	.43	.43	672	.78	.48	.48
M	4	1787	.78	.28	.28	675	.89	.32	.32
N	2	1782	.66	.44	.44	666	.76	.50	.50
N	3	1788	.83	.45	.45	670	.90	.40	.41

		Erhebungswelle 1				Erhebungswelle 7			
Itemblock	Itemnummer	Anzahl gültiger Werte	Schwierigkeit	Trennschärfe	Trennschärfe 47 Items	Anzahl gültiger Werte	Schwierigkeit	Trennschärfe	Trennschärfe 47 Items
O	1	1778	.70	.25	.26	669	.75	.31	.31
O	4	1783	.18	−.11		671	.20	.01	
P	2	1773	.63	.45	.45	672	.74	.47	.48
P	3	1794	.80	.37	.37	671	.87	.33	.33
Q	2	1794	.11	.07	.06	675	.12	.26	.25
Q	3	1780	.07	.04	.04	669	.08	.24	.22
R	2	1791	.10	.22	.21	673	.16	.41	.40
R	3	1772	.15	.19	.18	671	.18	.37	.36
S	3	1790	.07	.11	.10	669	.12	.29	.28
S	4	1774	.14	.06	.05	671	.22	.21	.20
T	1	1785	.82	.50	.50	672	.92	.45	.46
T	4	1791	.69	.41	.42	672	.77	.37	.38
U	1	1780	.54	.44	.44	671	.70	.48	.49
U	4	1779	.12	.07	.06	668	.16	.11	.10

Anmerkungen. Die Schwierigkeit entspricht dem Anteil korrekter Antworten. Die Trennschärfe entspricht der Korrelation des Items mit dem um das Item korrigierten Summenwert. Die Trennschärfe wurde einmal für alle 48 getesteten Items berechnet und einmal unter Ausschluss des Items Nr. 4 in Block O.

Tabelle A 3. Schwierigkeiten und Trennschärfen der Items des DGCF-MAT in der ersten Erhebungswelle

Setnummer	Itemnummer	Anzahl gültiger Werte	Schwierigkeit	Trennschärfe
1	1	1488	.75	.36
1	2	1490	.82	.31
1	3	1468	.31	.16
1	4	1390	.40	.28
2	1	1488	.57	.33
2	2	1455	.54	.43
2	3	1466	.32	.39
2	4	1461	.46	.45

Setnummer	Itemnummer	Anzahl gültiger Werte	Schwierigkeit	Trennschärfe
3	1	1491	.80	.36
3	2	1462	.61	.43
3	3	1478	.63	.40
3	4	1451	.35	.38

Anmerkungen. Die Schwierigkeit entspricht dem Anteil korrekter Antworten. Die Trennschärfe entspricht der Korrelation des Items mit dem um das Item korrigierten Summenwert.

Tabelle A 4. Schwierigkeiten und Trennschärfen der Can-Do-Statements zum Verstehen und Sprechen in beiden Erhebungswellen

	Erhebungswelle 1				Erhebungswelle 7			
Item-nummer	Anzahl gültiger Werte	Schwierigkeit	Trennschärfe	Trennschärfe 7 Items	Anzahl gültiger Werte	Schwierigkeit	Trennschärfe	Trennschärfe 7 Items
1	1870	.93	.46	.52	777	.94	.57	.61
2	1870	.90	.54	.59	777	.92	.64	.68
4	1870	.88	.56	.62	777	.91	.64	.67
6	1870	.85	.61	.65	777	.92	.67	.70
7	1870	.75	.54	.54	777	.86	.65	.67
10	1870	.39	.49	.44	777	.65	.53	.47
11	1870	.39	.46	.38	777	.64	.48	.42
14	1870	.29	.32		777	.61	.48	

Anmerkungen. Die Schwierigkeit entspricht dem Anteil korrekter Antworten. Die Trennschärfe entspricht der Korrelation des Items mit dem um das Item korrigierten Summenwert. Die Trennschärfe wurde einmal für alle 8 Can-Do-Statements zum Verstehen und Sprechen berechnet und einmal unter Ausschluss des Items Nr. 14.

Tabelle A 5. Schwierigkeiten und Trennschärfen der Items zur Nutzung von Möglichkeiten zum Deutschlernen in der ersten Erhebungswelle

Itemnummer	Anzahl gültiger Werte	Schwierigkeit	Trennschärfe	Trennschärfe 5 Items
1	1870	.36	.11	
2	1870	.36	.35	.36
3	1870	.70	.40	.70
4	1870	.55	.47	.55

Itemnummer	Anzahl gültiger Werte	Schwierigkeit	Trennschärfe	Trennschärfe 5 Items
5	1870	.63	.37	.63
6	1870	.34	.48	.34
7	1870	.62	.22	
8	1870	.16	.17	

Anmerkungen. Die Schwierigkeit entspricht dem Anteil korrekter Antworten. Die Trennschärfe entspricht der Korrelation des Items mit dem um das Item korrigierten Summenwert. Die Trennschärfe wurde einmal für alle 8 Items zur Nutzung von Möglichkeiten zum Deutschlernen berechnet und einmal unter Ausschluss der Items Nr. 1, 7 und 8.

Anhang B – Faktorladungen Skalen

Tabelle B 1. Rotierte Faktorladen der acht Can-Do-Statements zum Verstehen und Sprechen im zweifaktoriellen Modell in beiden Erhebungswellen

	Erhebungswelle 1		Erhebungswelle 7	
Itemnummer	**Faktor 1**	**Faktor 2**	**Faktor 1**	**Faktor 2**
1	–0.16	0.99	–0.19	1.07
2	0.01	0.93	0.01	0.94
4	–0.03	0.97	0.11	0.87
6	0.15	0.89	0.09	0.91
7	0.30	0.65	0.14	0.84
10	0.53	0.46	0.39	0.50
11	0.72	0.19	0.43	0.38
14	0.89	–0.13	0.96	0.02

Tabelle B 2. Rotierte Faktorladungen der Items zur Nutzung von Möglichkeiten zum Deutschlernen im dreifaktoriellen Modell in der ersten Erhebungswelle

Itemnummer	**Faktor 1**	**Faktor 2**	**Faktor 3**
1	0.02	0.57	0.03
2	–0.10	0.30	0.58
3	–0.00	0.14	0.65
4	–0.01	0.05	0.78
5	–0.03	–0.21	0.75
6	0.19	–0.03	0.73
7	0.97	–0.01	0.01
8	0.46	0.22	–0.00

Anhang C – Korrelationstabellen

Tabelle C 1. Korrelationen zwischen allen in den Analysen verwendeten Variablen der ersten Erhebungswelle

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1 Standard Verstehen	1 877	1 877	1 877	944	944	922	924	1 870	1 391	1 449	1 801	1 877	1 693	233	1 596	1 544	1 520	1 870	1 868	1 865
2 Standard Sprechen	.79	1 877	1 877	944	944	922	924	1 870	1 391	1 449	1 801	1 877	1 693	233	1 596	1 544	1 520	1 870	1 868	1 865
3 Standard Verstehen und Sprechen Mittelwert	.94	.95	1 877	944	944	922	924	1 870	1 391	1 449	1 801	1 877	1 693	233	1 596	1 544	1 520	1 870	1 868	1 865
4 Schieberegler Verstehen	.59	.58	.62	944	944	0	0	941	721	749	907	944	843	119	816	791	780	941	940	938
5 Schieberegler Verstehen und Sprechen Mittelwert	.61	.64	.66	.95	944	0	0	941	721	749	907	944	843	119	816	791	780	941	940	938
6 Vergleich Verstehen	.48	.49	.52	-	-	922	922	921	659	689	883	922	841	113	770	743	730	920	917	917
7 Vergleich Verstehen und Sprechen Mittelwert	.49	.52	.54	-	-	.96	924	923	661	691	885	924	843	114	772	745	732	922	919	919
8 Can-Do-Statements Verstehen und Sprechen Summenwert (7 Items)	.21	.20	.22	.26	.27	.25	.25	1 870	1 385	1 442	1 794	1 870	1 687	233	1 590	1 538	1 514	1 867	1 862	1 860
9 PPVT-4 Summenwert	.31	.30	.32	.32	.32	.33	.34	.23	1 391	1 391	1 330	1 391	1 238	183	1 349	1 332	1 331	1 384	1 385	1 381
10 PPVT-4 WLEs	.32	.31	.34	.33	.33	.35	.35	.24	.99	1 449	1 373	1 449	1 292	190	1 405	1 388	1 386	1 442	1 443	1 439
11 TROG-D Summenwert (47 Items)	.28	.28	.30	.28	.30	.29	.29	.25	.70	.72	1 801	1 801	1 620	224	1 533	1 484	1 458	1 794	1 793	1 789
12 Kombinierter Deutschkompetenzscore	.30	.29	.31	.31	.32	.30	.31	.25	.91	.93	.95	1 877	1 693	233	1 596	1 544	1 520	1 870	1 868	1 865

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
13 Leistung in Mathematik (Selbstbericht)	.19	.17	.19	.26	.26	.08	.09	.13	.13	.12	.12	.13	1 693	229	1 433	1 384	1 362	1 687	1 684	1 683
14 Leistung in Mathematik (Angabe Lehrkraft)	−.00	.02	.01	.19	.23	−.10	−.08	.07	.37	.37	.28	.33	.52	233	206	197	197	233	232	232
15 DGCF-MAT Set 1 Summenwert	.06	.02	.04	.06	.04	.06	.06	.11	.20	.19	.22	.22	.12	.09	1 596	1 543	1 519	1 590	1 590	1 586
16 DGCF-MAT Set 2 Summenwert	.08	.06	.07	.10	.09	.06	.06	.10	.28	.27	.27	.28	.20	.24	.40	1 544	1 504	1 538	1 538	1 534
17 DGCF-MAT Set 3 Summenwert	.08	.06	.07	.07	.06	.05	.05	.09	.23	.23	.27	.27	.14	.23	.39	.52	1 520	1 514	1 514	1 510
18 Nutzung Möglichkeiten Deutschlernen Summenwert (5 Items)	.12	.11	.12	.17	.17	.10	.12	.33	.17	.16	.12	.13	.15	.09	.07	.13	.12	1 870	1 861	1 861
19 Teilnahme Deutschkurs	.17	.16	.18	.15	.18	.20	.20	.13	.00	.01	.03	.01	.09	.15	.00	.00	.03	.07	1 868	1 860
20 Teilnahme Deutschtest	.15	.12	.14	.11	.11	.16	.15	.07	.03	.05	.10	.07	.06	.03	.03	−.05	.02	−.06	.47	1 865

Anmerkungen. Pearson-Korrelationen unterhalb der Diagonalen, Stichprobengrößen der paarweisen Korrelationen oberhalb der Diagonalen. WLE = Weighted Likelihood Estimate.

Tabelle C 2. Korrelationen zwischen allen in den Analysen verwendeten Variablen der siebten Erhebungswelle

	1	2	3	4	5	6
1 Standard Verstehen	778	777	777	773	744	778
2 Standard Sprechen	.60	777	776	772	743	777

	1	2	3	4	5	6
3 Can-Do-Statements Verstehen und Sprechen Summenwert (7 Items)	.32	.26	777	772	743	777
4 PPVT-4 Summenwert	.33	.27	.36	773	739	773
5 TROG-D Summenwert (47 Items)	.35	.30	.37	.71	744	744
6 Kombinierter Deutschkompetenzscore	.37	.30	.39	.93	.93	778

Anmerkungen. Pearson-Korrelationen unterhalb der Diagonalen, Stichprobengrößen der paarweisen Korrelationen oberhalb der Diagonalen.

Anhang D – Korrelationen der Residuen in den Strukturgleichungsmodellen

Tabelle D 1. Korrelationen der Residuen der Indikatoren im Messmodell zu den Einflussfaktoren der Selbsteinschätzungen (Modellierungsvariante 1)

	1	2	3	4	5	6	7	8	9	10
1 PPVT-4 WLE										
2 TROG-D Summenwert (47 Items)	.00									
3 Standard Verstehen und Sprechen Mittelwert	.01	.00								
4 Schieberegler Verstehen und Sprechen Mittelwert	–.02	–.03	.00							
5 Vergleich Verstehen und Sprechen Mittelwert	.07	.03	.00	**–.28**						
6 Leistung in Mathematik (Selbstbericht)	–.01	–.02	.01	.06	–.07					
7 Leistung in Mathematik (Angabe der Lehrkraft)	**.24**	**.16**	**–.10**	.07	**–.11**	–.02				
8 Nutzung von Möglichkeiten Deutschlernen Summenwert (5 Items)	.00	–.01	–.01	.01	.01	.00	–.05			
9 DGCF-MAT Set 1 Summenwert	.00	.02	–.01	.01	.01	–.01	–.01	–.02		
10 DGCF-MAT Set 2 Summenwert	.00	.00	.00	.01	.01	.02	**.12**	.01	–.01	
11 DGCF-MAT Set 3 Summenwert	–.03	.02	.00	–.01	.00	–.03	**.11**	.00	.01	.00

Anmerkungen. Korrelationen > |.10| sind hervorgehoben. WLE = Weighted Likelihood Estimate.

Tabelle D 2. Korrelationen der Residuen der Indikatoren im Messmodell zu den Einflussfaktoren der Selbsteinschätzungen (Modellierungsvariante 2)

	1	2	3	4	5	6	7	8	9	10
1 PPVT-4 WLE										
2 TROG-D Summenwert (47 Items)	.00									
3 Standard Verstehen und Sprechen Mittelwert	.01	–.01								
4 Schieberegler Verstehen und Sprechen Mittelwert	.00	.00	–.01							
5 Vergleich Verstehen und Sprechen Mittelwert	.03	–.01	.01	**–.31**						
6 Leistung in Mathematik	–.01	–.01	.01	.05	–.07					

(Selbstbericht)											
7 Leistung in Mathematik (Angabe der Lehrkraft)	**.23**	**.16**	**–.10**	.06	**–.12**	–.02					
8 Nutzung von Möglichkeiten Deutschlernen Summenwert (5 Items)	.00	.00	–.01	.01	.00	.00	–.05				
9 DGCF-MAT Set 1 Summenwert	–.01	.02	–.01	.02	.00	–.01	–.01	–.02			
10 DGCF-MAT Set 2 Summenwert	.00	.00	.00	.02	–.01	.02	**.12**	.01	–.01		
11 DGCF-MAT Set 3 Summenwert	–.03	.02	.00	.00	–.02	–.03	.10	.00	.01	.00	

Anmerkungen. Korrelationen > |.10| sind hervorgehoben. WLE = Weighted Likelihood Estimate.

Tabelle D 3. Korrelationen der Residuen im Gesamtmodell zu den Einflussfaktoren der Selbsteinschätzungen (Modellierungsvariante 1)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1 PPVT-4 WLE												
2 TROG-D Summenwert (47 Items)	.00											
3 Standard Verstehen und Sprechen Mittelwert	.01	.00										
4 Schieberegler Verstehen und Sprechen Mittelwert	–.02	–.02	.01									
5 Vergleich Verstehen und Sprechen Mittelwert	.06	.03	.00	**–.28**								
6 Leistung in Mathematik (Selbstbericht)	–.01	–.01	.01	.06	–.07							
7 Leistung in Mathematik (Angabe der Lehrkraft)	**.22**	**.16**	–.10	.07	**–.12**	–.03						
8 Nutzung von Möglichkeiten Deutschlernen Summenwert (5 Items)	.00	–.01	–.01	.01	.01	.00	–.06					
9 DGCF-MAT Set 1 Summenwert	–.01	.02	–.01	.01	.01	–.01	–.02	–.02				
10 DGCF-MAT Set 2 Summenwert	–.01	.00	.00	.01	.01	.02	.10	.01	–.01			
11 DGCF-MAT Set 3 Summenwert	–.03	.02	.00	–.01	.00	–.03	.09	.00	.01	.00		
12 Teilnahme Deutschkurs	.00	.03	.00	–.01	.03	.00	.01	.00	–.01	.00	.02	
13 Teilnahme Deutschunterricht	.04	.10	.03	.00	.05	.01	–.01	.01	.05	–.02	.05	.00

Anmerkungen. Korrelationen > |.10| sind hervorgehoben. WLE = Weighted Likelihood Estimate.

Tabelle D 4. Korrelationen der Residuen im Gesamtmodell zu den Einflussfaktoren der Selbsteinschätzungen (Modellierungsvariante 2)

	1	2	3	4	5	6	7	8	9	10	11	12
1 PPVT-4 WLE												
2 TROG-D Summenwert (47 Items)	.00											
3 Standard Verstehen und Sprechen Mittelwert	.01	.00										
4 Schieberegler Verstehen und Sprechen Mittelwert	.00	.00	.00									
5 Vergleich Verstehen und Sprechen Mittelwert	.03	–.01	.00	**–.29**								
6 Leistung in Mathematik (Selbstbericht)	–.01	–.01	.01	.06	–.07							
7 Leistung in Mathematik (Angabe der Lehrkraft)	**.22**	**.16**	–.10	.06	**–.12**	–.03						
8 Nutzung von Möglichkeiten Deutschlernen Summenwert (5 Items)	.00	–.01	–.01	.01	.01	.00	–.06					
9 DGCF-MAT Set 1 Summenwert	–.01	.02	–.01	.02	.00	–.01	–.02	–.02				
10 DGCF-MAT Set 2 Summenwert	–.01	.00	.00	.02	.00	.02	.10	.01	–.01			
11 DGCF-MAT Set 3 Summenwert	–.03	.02	.00	.00	–.02	–.03	.09	.00	.01	.00		
12 Teilnahme Deutschkurs	.00	.03	.01	–.02	.04	.00	.01	.00	–.01	.00	.02	
13 Teilnahme Deutschunterricht	.04	.10	.04	–.01	.05	.01	–.01	.01	.05	–.02	.05	.00

Anmerkungen. Korrelationen > |.10| sind hervorgehoben. WLE = Weighted Likelihood Estimate.

Tabelle D 5. Korrelationen der Residuen im Messmodell starker faktorieller Invarianz zur Veränderung der Kompetenzmaße über die Zeit

	1	2	3	4	5	6	7	8	9
1 PPVT-4 Summenwert zentriert 1. Erhebungswelle									
2 TROG-D Summenwert (47 Items) zentriert 1. Erhebungswelle	.01								
3 PPVT-4 Summenwert zentriert 7. Erhebungswelle	.04	.02							
4 TROG-D Summenwert (47 Items) zentriert 7. Erhebungswelle	–.10	.02	.00						
5 Standard Verstehen 1. Erhebungswelle	.01	–.03	–.02	.00					
6 Standard Sprechen 1. Erhebungswelle	.03	.02	.01	.03	.00				
7 Standard Verstehen 7. Erhebungswelle	.00	.00	.02	.03	–.02	–.01			

	1	2	3	4	5	6	7	8	9
8 Standard Sprechen 7. Erhebungswelle	–.01	.01	–.05	–.02	.00	.06	.00		
9 Can-Do-Statements Verstehen und Sprechen Summenwert (7 Items) 1. Erhebungswelle	–.04	.03	–.02	.03	.00	.01	–.01	.02	
10 Can-Do-Statements Verstehen und Sprechen Summenwert (7 Items) 7. Erhebungswelle	–.05	.03	–.01	.01	.00	.01	.02	–.03	.00

Tabelle D 6. Korrelationen der Residuen im Gesamtmodell zur Veränderung der Kompetenzmaße über die Zeit

	1	2	3	4	5	6	7	8	9
1 PPVT-4 Summenwert zentriert 1. Erhebungswelle									
2 TROG-D Summenwert (47 Items) zentriert 1. Erhebungswelle	.01								
3 PPVT-4 Summenwert zentriert 7. Erhebungswelle	.04	.02							
4 TROG-D Summenwert (47 Items) zentriert 7. Erhebungswelle	–.10	.02	.00						
5 Standard Verstehen 1. Erhebungswelle	.01	–.03	–.02	.00					
6 Standard Sprechen 1. Erhebungswelle	.03	.02	.01	.03	.00				
7 Standard Verstehen 7. Erhebungswelle	.00	.00	.02	.03	–.02	–.01			
8 Standard Sprechen 7. Erhebungswelle	–.01	.01	–.05	–.02	.00	.06	.00		
9 Can-Do-Statements Verstehen und Sprechen Summenwert (7 Items) 1. Erhebungswelle	–.04	.03	–.02	.03	.00	.01	–.01	.02	
10 Can-Do-Statements Verstehen und Sprechen Summenwert (7 Items) 7. Erhebungswelle	–.05	.03	–.01	.01	.00	.01	.02	–.03	.00