Zeitschrift: Bulletin suisse de linguistique appliquée / VALS-ASLA

Band: - (2021)

Heft: 113: Wortschatzkompetenzen definieren, erheben und fördern = Defining, assessing and fostering vocabulary skills

Artikel: Bedeutung und Diagnostik des Wortschatzes am Beispiel des Peabody Picture Vocabulary Test (PPVT-IV)

Autor: Lenhard, Wolfgang / Lenhard, Alexandra

DOI: https://doi.org/10.5169/seals-1030123

Bedeutung und Diagnostik des Wortschatzes am Beispiel des Peabody Picture Vocabulary Test (PPVT-IV)

Wolfgang LENHARD
Universität Würzburg
Institut für Psychologie
Wittelsbacherplatz 1, D-97074 Würzburg
wolfgang.lenhard@uni-wuerzburg.de

Alexandra LENHARD
Psychometrica
Am Kreuz 14, D-97337 Dettelbach
lenhard@psychometrica.de

Vocabulary is a fundamental determinant of language development. In the Cattell-Horn-Carroll model of intelligence (CHC) it represents a narrow ability loading on the broad comprehension-knowledge (Gc) factor of intelligence. Furthermore, it is also an important predictor of reading comprehension. The Peabody Picture Vocabulary Test (Version 4) aims at assessing receptive vocabulary. Here, we describe the German adaption of the test. In the construction, we matched word frequencies and item complexity of the original form. The normative data is based on a representative sample of N = 3550 children and adolescents with an age range from 2.59 to 17.99 years. The test features excellent reliability. The raw scores display a strong curvilinear development during childhood and a tremendous heterogeneity within each age group. The best performing children at the age of 5 achieved raw scores which the poor performing children did not achieve until the age of 15. While effects of sex were negligible, migration background of the family had a strong effect. In sum, the PPVT-IV is an economic and reliable instrument to assess the receptive vocabulary from an early age until adulthood.

Keywords:
vocabulary assessment, bilingualism, cognitive abilities, word knowledge, language development.

Stichwörter:
Wortschatztest, Bilingualismus, kognitive Fähigkeiten, Wortschatzwissen, Sprachentwicklung.

1. Die Bedeutung des Wortschatzes für menschliche Kognition

Der Begriff "Wortschatz" ist sehr schillernd, da er selbst auf seine enorme Bedeutung verweist. Ein Schatz ist etwas Kostbares, das man hegt und pflegt und zu vergrößern trachtet. Während materielle Schätze auch schlicht gefunden oder geerbt werden können, verhält es sich mit dem Wortschatz auf individueller Ebene anders. Der Erwerb des Wortschatzes stellt eine Form von Wissenserwerb dar, welcher sich über einen langen Zeitraum hinzieht und erst im mittleren Erwachsenenalter den Höhepunkt erreicht (Wechsler 2008). Er findet entweder vorwiegend implizit im Rahmen des Erstspracherwerbs oder explizit im Rahmen von formellem (Zweit- oder Fremd-)Sprachunterricht statt. Welche spezifischen Wörter gelernt werden, hängt nicht nur von der erlernten Sprache, sondern auch vom Kontext des Erwerbs ab, also davon, ob die Wörter

in häuslichen, akademischen, beruflichen oder anderen Situationen erworben werden. Wortbedeutungen können sich dabei mit der Zeit nicht nur gesamtgesellschaftlich, sondern auch auf individueller Ebene verändern, d. h. sie werden durch neue Bedeutungsfacetten bereichert, verändern ihren semantischen Gehalt, werden von anderen Wörtern verdrängt und möglicherweise zu einem späteren Zeitpunkt erneut reaktiviert. Auf individueller Ebene übersteigt der Wortschatz einer Sprache das Wortschatzwissen der einzelnen Personen deutlich. Anders als andere Fähigkeits- und Wissensbereiche wie beispielsweise das Phoneminventar einer Sprache kann der Wortschatz also fortlaufend erweitert und vertieft werden und ist somit ein lohnenswertes Ziel für jegliche Form von Bildungsprozessen.

1.1 Wortschatz und Intelligenz

Die Bedeutung des Wortschatzes liegt nicht alleine im Umfang des verfügbaren Vokabulars, sondern auch in seiner Tiefe, also der Reichhaltigkeit der verfügbaren Wortbedeutungen, der Flexibilität der Anwendung und den Querverbindungen oder Relationen zwischen sprachlichen Konzepten. Wissen organisiert sich meist um eine spezifische Terminologie, bzw. beinhaltet diese, sodass sich i. d. R. hohe Korrelationen zwischen bereichsspezifischem Wissen und Wortschatz ergeben (z. B. Cromley & Azvedo 2007). Auch beeinflusst die Größe des Wortschatzes maßgeblich das Ziehen von Schlussfolgerungen beim Lesen von Sachtexten und damit das gesamte Textverständnis.

Die Bedeutung des Wortschatzes geht also weit über eine rein sprachliche Ebene hinaus und kann als ein konstitutives Merkmal menschlicher Kognition angesehen werden. Die Stellung des Wortschatzes innerhalb des Spektrums menschlicher Intelligenzleistungen ist durch weithin anerkannte faktorenanalytische Modelle der Intelligenz empirisch hervorragend belegt (Schneider & McGrew 2018; vgl. Abbildung 1). So enthalten viele wichtige Intelligenztests (z. B. die Wechsler-Tests und die Kaufman-Tests) auch Untertests zur Erfassung des Wortschatzes. Der Anteil sprachlicher Inhalte an Intelligenztests wurde im Laufe der Jahrzehnte tendenziell sogar aufgewertet, da sich der Umgang mit sprachlichem Material als ein fast unverzichtbarer Prädiktor zur Vorhersage der allgemeinen Intelligenz einer Person herausstellte. Testverfahren, die auch sprachliche Aufgaben beinhalten, sagen schulische Leistungen besser vorher als rein non-verbale Tests (Roth et al. 2015; A. Lenhard 2020). Das gilt nicht alleine für sprachliche Fächer, sondern auch für das Schulfach Mathematik.

Eines der aktuell etabliertesten Intelligenzmodelle ist das Cattell-Horn-Carroll-Modell (CHC; z. B. Schneider & McGrew 2018), das auf Grundlage hunderter empirischer Primärstudien konzipiert wurde. Das Modell listet über 80 sog. engere Intelligenzfähigkeiten auf, die anhand des Kovarianzmusters zu breiteren Intelligenzfaktoren gruppiert werden. Die gemeinsame Varianz aller Intelligenzleistungen wird als allgemeine Intelligenz G bezeichnet. Unter den

breiten Faktoren finden sich beispielsweise die *Arbeitsgedächtniskapazität* G_{wm}, die *Verarbeitsgeschwindigkeit* G_s, die sogenannte *fluide Intelligenz* G_f (d. h. die Fähigkeit zum Erkennen und Anwenden von Mustern und Gesetzmäßigkeiten), aber auch das *Sprachverständnis und -wissen* (G_c). Der letztgenannte Faktor wird dabei durch die engen Fähigkeiten *Sprachentwicklung*, *allgemeines Wissen*, *Hörverständnis*, *Kommunikationsfähigkeit*, *Grammatikempfinden*, aber eben auch *lexikalisches Wissen* – sprich: Wortschatz – aufgespannt. Die Bedeutung der engen Fähigkeiten für die breiten Faktoren lässt sich in den statistischen Modellen über die Faktorladungen spezifizieren, mit denen die Einzelleistungen auf dem übergeordneten latenten Konstrukt laden. Betrachtet man die bestehenden Modelle, so zeigt sich, dass Wortschatzleistungen unter den verschiedenen Einzelleistungen die höchsten Ladungen auf G_c aufweisen und dieser Faktor lädt wiederum hoch auf der allgemeinen Intelligenz (z. B. im Testverfahren WISC-V; Wechsler 2014). Der Wortschatz ist somit nicht nur einer der besten Indikatoren für das Sprachverständnis, sondern auch für die allgemeine Intelligenz selbst. In Large-Scale-Studien wird der rezeptive Wortschatz deshalb zuweilen sogar als alleiniger Indikator für G_c, manchmal sogar auch als Schätzer für die allgemeine Intelligenz G erhoben (z. B. British Cohort Study, Bynner & Parsons 2005). Auch Carroll (1993) zog in seinen Arbeiten den Schluss, dass Wortschatz eine sehr hohe Überlappung mit der breiteren Dimension "verbale Intelligenz" aufweist und unterstrich damit die Bedeutung des Wortschatzes für die kognitiven Fähigkeiten einer Person.

Wortschatz und Intelligenz hängen vermutlich auf vielfältige Weise miteinander zusammen und bedingen sich gegenseitig (vgl. A. Lenhard et al. 2015). Beispielsweise können Personen mit höherer fluider Intelligenz die Bedeutung von Wörtern besser aus dem Kontext erschließen und implizit einen größeren Wortschatz erwerben. Auf der anderen Seite erleichtert ein besser vernetzter Wortschatz vermutlich wiederum das Erkennen von sprachlichen Konzepten und Zusammenhängen, wie die Identifikation von Gemeinsamkeiten und Bedeutungsunterschieden, d. h. das logische Denken mit Sprache. Die Herausbildung von Unter- und Oberbegriffen wird unterstützt und neues Wissen besser organisiert und schneller erworben. Ein großer Wortschatz erleichtert zudem viele andere akademische Tätigkeiten, sodass Personen mit einem größeren Wortschatz tendenziell mehr lesen und beim Lesen wiederum mehr neuen Wortschatz erwerben (Pfost et al. 2013). Sie verfügen über einen größeren Umfang an sprachlichen Konzepten, auf die sie als Werkzeug beim Denken zugreifen können. Es lässt sich somit festhalten, dass der Wortschatz zwar einen abgrenzbaren Bereich innerhalb der Palette kognitiver Fähigkeiten und Fertigkeiten aufspannt, aber mit vielen weiteren Intelligenzfacetten interagiert und deshalb zugleich Ursache und Wirkung der Intelligenzentwicklung darstellt.

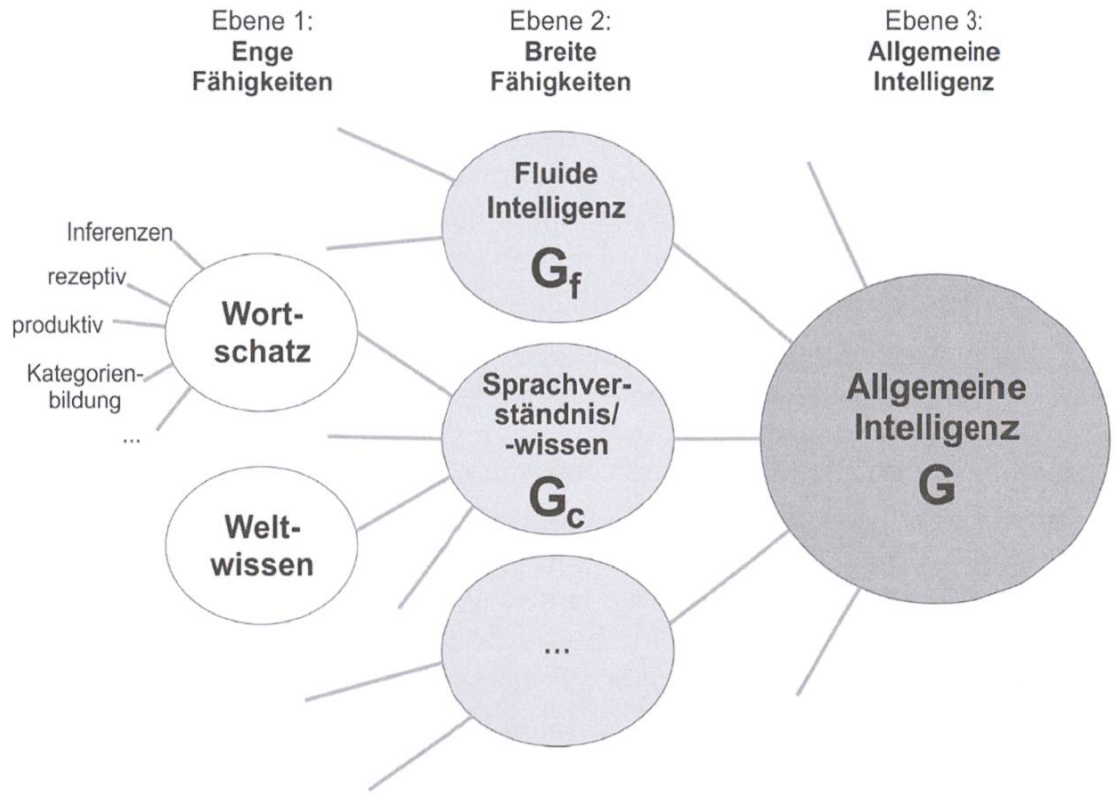

Abbildung 1: Das CHC-Rahmenmodell der Intelligenz. Das Cattell-Horn-Carroll-Modell (CHC; z. B. Schneider & McGrew 2018) definiert Intelligenz als hierarchische Struktur mit eng umrissenen kognitiven Fähigkeiten, die sich auf Ebene 2 des Modells zu breiten Faktoren gruppieren. Die allgemeine Intelligenz ist in diesem Modell die gemeinsame Varianz, die sich aus den Intelligenzfaktoren extrahieren lässt. Manche Aufgaben, wie z. B. das Auffinden von Gemeinsamkeiten oder Unterschieden zwischen Begriffen, hängen sowohl mit G_c als auch mit G_f eng zusammen.

1.2 Wortschatz, Schriftsprache und schulischer Wissenserwerb

Je jünger Kinder sind, desto weniger formell erfolgt in der Regel der Wortschatzerwerb. Doch bereits im ersten Lebensjahr werden die Grundlagen für die spätere Größe des Wortschatzes gelegt. So wirkt sich gemeinsames Bücherlesen vermutlich bereits Ende des ersten Lebensjahres förderlich aus (Lenhart et al., im Druck). Zunächst besteht dieses gemeinsame "Lesen" natürlich v. a. im Benennen und Zeigen von Objekten in Bilderbüchern, später dann im (dialogischen) Vorlesen oder dem freien Geschichtenerzählen. Je höher der Grad an Interaktion ist, desto mehr neuer Wortschatz wird erworben. In eigenen Untersuchungen erwies sich insbesondere das freie Geschichtenerzählen im Vorschulalter dem Vorlesen von Geschichten als überlegen, da es dabei ein größeres Ausmaß an Interaktionen gibt (Suggate et al. 2021). Der Lerneffekt kann noch gesteigert werden, indem Kinder die Geschichten nicht nur hören, sondern anschließend selbst nacherzählen ("Interactive Elaborative Story Retelling"; Vaahtoranta et al. 2019).

Ein sehr starker Zusammenhang besteht natürlich nicht nur zwischen Wortschatz und allgemeiner Intelligenz, sondern auch spezifisch zwischen Wortschatz und Leseverständnis (Beck et al. 1982; Verhoeven et al. 2011). Zusammen mit anderen Intelligenzleistungen wie z. B. Arbeitsgedächtniskapazität und grammatikalischen Fähigkeiten ist die Kenntnis von Wortbedeutungen von zentraler Bedeutung beim verstehenden Lesen von Texten. Cromley und Azevedo (2007) quantifizierten die direkten und indirekten Einflüsse verschiedener Determinanten des Leseverständnisses, wie z. B. Wortschatz, Leseflüssigkeit, Hintergrundwissen, Strategiewissen und schlussfolgerndes Denken. Der Wortschatz erwies sich in dieser Untersuchung als der mit weitem Abstand bedeutendste Einflussfaktor. Wie bereits dargestellt, erfassen Menschen die Bedeutung von neuen Wörtern unter anderem, indem sie den Sinn unbekannter Wörter, denen sie im Text begegnen, aufgrund des Kontexts erschließen (siehe Beitrag von Brugger & Juska-Bacher 2021). Allerdings ist es für Personen mit einem geringen Wortschatz oftmals überproportional anstrengend, Texte zu verstehen, da der Verständnisprozess umso brüchiger wird, je größer die Verständnislücken sind. In der Folge kann sich ein Teufelskreis aus Lesevermeidung, mangelnder Automatisierung, und geringerem Wissens- und Wortschatzerwerb entwickeln, durch den ein weiterer Wortschatzaufbau verzögert und die Leseentwicklung weiter unterbunden wird.

Lesefähigkeiten werden jedoch nicht automatisch erworben, indem man viel liest (Hattie 2010: 140). So führt außerschulisches Lesen v. a. bei bildungsnahen Schichten zu einer Zunahme der Lesekompetenz (Pfost et al. 2010). Ähnlich verhält es sich mit dem Erwerb neuer Wörter: Die Lesekompetenz muss überhaupt erst hinreichend gut ausgeprägt sein, damit Kinder neuen Wortschatz implizit beim Lesen erlernen können. In einer eigenen Untersuchung erwarben beispielsweise Kinder im Grundschulalter zunächst die Bedeutung unbekannter Wörter v. a. beim Geschichtenerzählen, aber nicht, wenn sie die Texte selbst lasen (Suggate et al. 2013). Erst wenn Kinder sich am Ende der Grundschulzeit befanden, gelang der implizite Erwerb von Wortschatz ähnlich gut wie beim Geschichtenerzählen.

2. Spektrum der Wortschatzdiagnostik

Wortschatzdiagnostik war bereits in den Frühphasen der Psychometrie Gegenstand der Forschung (z. B. Sims 1929). Die Erfassung des Wortschatzes kann nicht nur nach Aufgabenformen, sondern auch danach gruppiert werden, wie tief das erfasste Wissen reicht, nämlich vom bloßen Wiedererkennen der phonetischen oder visuellen Gestalt bis zum situationsangemessenen Anwenden und der Fähigkeit, das sprachliche Konzept zu erläutern und Querverbindungen zu anderen Wörtern zu ziehen. Im Folgenden werden die verschiedenen Ansätze charakterisiert und im Hinblick auf ihre Vor- und Nachteile diskutiert.

Auf der Ebene des bloßen Wiedererkennens von Wörtern entstanden in den letzten Jahren verschiedene Variationen der sog. *Lexical Decision Tasks*. Wörter werden dabei entweder akustisch oder visuell präsentiert. Die Testperson muss entscheiden, ob es sich dabei um ein reales Wort handelt oder nicht, bzw. ob die Person das Wort kennt (z. B. Lemhöfer & Broersma 2012 und die darauf aufbauende Online-Plattform http://www.lextale.com/; Trautwein & Schroeder 2018). Solche Tests sind einfach zu konstruieren und sie weisen eine hohe Ökonomie auf. Es existieren sogar Ansätze zur automatisierten Generierung von falschen Antwortalternativen (Hamed & Zesch 2018). Denn um reales Wissen von einer reinen Ja-Sage-Tendenz abzugrenzen, müssen zu solchen Tests auch Pseudowörter beigemischt werden. Gerade hierbei ergibt sich jedoch eine Reihe an Fallstricken. Aus methodischer Sicht stellt sich nämlich nicht nur die Frage, wie man die korrekten Wörter auswählt und nach welchen Kriterien das Aufgabenmaterial ausbalanciert wird, sondern auch, wie man Pseudowörter generiert, die realen Wörtern hinreichend ähnlich sind. Beispielsweise ist es möglich, einfach Anfangsbuchstaben zu ersetzen. Bei der schriftlichen Darbietung liegt allerdings eine Gefahr darin, Pseudohomophone zu generieren, also Wörter, deren lautliche Realisation einem realen Wort entspricht, das aber lediglich falsch geschrieben ist. Berücksichtigt man solche Probleme nicht, dann wird eher orthographisches Wissen erfasst. Dies kann durchaus auch das Ziel einer Untersuchung sein, aber man sollte sich der Konfundierung bewusst sein. Zum anderen stellen solche Tests auch eine sehr oberflächliche Überprüfung des Wortschatzes dar, da sie das bloße Wiedererkennen der Wortgestalt ohne irgendeine Form tiefergehenden lexikalischen Wissens bzw. Kenntnis der Semantik erfassen. Auch eine reine Stichwortsuchmaschine würde einen solchen Test also mit Bestergebnissen bestehen.

Validere Ergebnisse erzielt man deshalb in der Regel mit Tests, bei denen eine Form des semantischen Verständnisses erfasst wird. Dies geschieht beispielsweise mit Wort-Bild-Zuordnungen (z. B. der Peabody Picture Vocabulary Test 4: PPVT-IV; deutschsprachige Adaptation: A. Lenhard et al. 2015; s. u.) oder indem Synonyme, Antonyme oder Unter- und Oberbegriffe identifiziert werden müssen. Letzteres wird häufig in Intelligenztestbatterien eingesetzt (z. B. BIS-HB, Jäger et al. 2006; KFT 4-12+, Heller & Perleth 2000). Die erfassten Fähigkeiten reichen hier also von einer bloßen Wiedererkennung der Wortbedeutung bis zu einem tieferen Verständnis der Bezüge zwischen Wörtern. Zwar liefern auch diese Tests noch keinen Nachweis dafür, dass eine Testperson die Wörter auch kontextangemessen einsetzen kann. Außerdem ist der Aufwand bei der Testkonstruktion generell höher als bei lexical decision tasks. Der große Vorteil liegt jedoch in der hohen Reliabilität und Objektivität der geschlossenen Aufgabenformate. Zudem kann auch hier angesichts der hohen Ökonomie in der Anwendung ein breites Wortschatzwissen abgeprüft werden.

Und schließlich kann auch der produktive Wortschatzgebrauch überprüft werden. Dies wird beispielsweise mit Lückentexten wie den C-Tests gemacht, die v. a. im Rahmen der Fremdsprachdiagnostik eingesetzt werden (methodisch Diskussion siehe Eckes & Grotjahn 2006; Khodadady 2014). Allerdings werden hierbei über das Wortschatzwissen hinaus auch syntaktische (und bei automatisierter Darbietung orthografische) Fähigkeiten miterfasst.

Besonders tiefgehend wird das Verständnis für die Konzepte hinter den Wörtern in Tests erfasst, bei denen Personen die Wortbedeutung in eigenen Worten erklären müssen. Dies ist beispielsweise bei Untertests der Skala *Sprachverständnis* der Wechsler-Reihe (z. B. WAIS-IV, Wechsler 2008; WISC-V, Wechsler 2014) der Fall. Beim Untertest *Gemeinsamkeiten finden* müssen Personen beispielsweise Vergleiche zwischen verschiedenen, teilweise sehr abstrakten sprachlichen Konstrukten ziehen, beim *Wortschatz-Test* müssen sie die zentralen Aspekte eines sprachlichen Konzepts erklären. Die abgefragten Bedeutungsfacetten sind dabei z. T. sehr komplex und abstrakt. Die erfassten Fähigkeiten gehen deshalb stark in Richtung allgemeine Intelligenz. Eine Schwierigkeit hierbei besteht außerdem in der Bewertung der Aufgabenlösungen, da diese nicht ausschließlich mit richtig vs. falsch bewertet werden können. Bei der Testauswertung besteht deshalb ein größerer subjektiver Bewertungsspielraum als bei einfacheren Wortschatztests.

3. Peabody Picture Vocabulary Test IV (PPVT-IV)

Ein im internationalen Bereich sehr weit verbreitetes, aus dem englischen Sprachraum stammendes, standardisiertes Testverfahren liegt mit dem PPVT vor, der aktuell in der vierten Ausgabe verfügbar ist. Das Verfahren geht ursprünglich auf das Ehepaar Lloyd und Leona Dunn zurück und hat seitdem eine Reihe an Überarbeitungen erfahren, an denen auch deren Sohn Douglas Dunn mitgewirkt hat. Das Verfahren wurde in zahlreiche Sprachen übertragen bzw. für diese anhand des ursprünglichen Konstruktionsprinzips neu entwickelt. Die Adaptionen sind so weit wie möglich zur Originalfassung parallelisiert, sodass sprachübergreifend ein Untersuchungsinstrument zur Erfassung des rezeptiven Wortschatzes zur Verfügung steht. Auch die deutschsprachige Adaption nahm ihren Ausgangspunkt in einer interkulturellen Vergleichsstudie zum Spracherwerb (Suggate et al. 2014), bei dem der vorschulische Spracherwerb in Deutschland und Neuseeland und die Bedeutung verschiedener Prädiktoren auf den schulischen Schriftspracherwerb längsschnittlich verglichen wurden. Die folgenden Ausführungen beziehen sich schwerpunktmäßig auf die deutschsprachige Adaption (Lenhard et al. 2015) und die Ergebnisse der zugehörigen Normierungsstudie.

3.1 Testlogik und Ablauf

Ziel des PPVT ist die Erfassung des rezeptiven Wortschatzes. Hierzu wird jeweils eine Bildkarte mit vier Bildern gezeigt und dazu ein Wort vorgesprochen. Die Testperson muss auf das Bild deuten, das am besten zum Wort passt. Insgesamt gibt es im Testverfahren 228 Wörter und zugehörige Bildkarten, die der Schwierigkeit nach angeordnet und in 12er-Blöcke gruppiert sind. Es existieren folglich insgesamt 19 dieser Aufgabenblöcke, die jeweils mit einem altersspezifischen Einstiegskriterium versehen sind. So beginnen beispielsweise Kinder im Alter von 5 Jahren bei Aufgabenset 4, Jugendliche ab 14 Jahren dagegen erst mit Aufgabenblock 11. Werden in diesem Block mindestens 11 der 12 Wörter korrekt erkannt, so gelten alle Aufgaben in niedrigeren Blöcken automatisch als bestanden. Ist dies nicht der Fall, so wird zunächst das Bodenset ermittelt, indem Block für Block zurückgegangen wird, bis schließlich ein Aufgabenblock mit mindestens 11 korrekten Lösungen gefunden wurde. Anschließend wird die Testung mit schwierigeren Aufgaben fortgesetzt, bis schließlich in einem Block acht oder mehr Fehler aufgetreten sind.

Wenn eine Testperson bei einer Aufgabe nicht innerhalb von etwa 10 Sekunden antwortet, so wird eine Ermutigung gegeben, z. B. "Versuche es einfach. Zeige auf dasjenige, von dem Du denkst, es könnte richtig sein". Wenn die Testperson immer noch nicht antwortet, wertet man das Item als Fehler, protokolliert die fehlende Antwort und geht zügig zur nächsten Aufgabe über. Die Gesamtpunktzahl ergibt sich schließlich aus der Anzahl richtig gelöster Aufgaben plus der Anzahl an automatisch als korrekt gewerteten Aufgaben unterhalb des Bodensets. Die Testdarbietung erfolgt ohne Zeitbegrenzung. Da Testpersonen unterschiedlich schnell arbeiten, hängt die Dauer der Untersuchung von der untersuchten Person ab. Im Schnitt dauert die Testung allerdings weniger als 15 Minuten, in nur 10% aller Durchführungen werden 20 Minuten oder mehr benötigt.

3.2 Auswahl und Adaption des Aufgabenmaterials

Die ursprüngliche Fassung des PPVT von 1953 hatte zunächst kein linguistisch orientiertes Konstruktionsmodell, sondern die Auswahl der Wörter geschah durch Sichtung aller Wörter in Standardlexika, die sich gut für eine Visualisierung eigneten. In die ursprünglichen Datenerhebungen gingen 600 Bildkarten mit 2400 Begriffen ein, aus denen schließlich 300 Aufgaben, aufgeteilt auf zwei Testfassungen, ausgewählt wurden. Fassung 2 und 3 zeichneten sich durch eine Verbesserung der psychometrischen Eigenschaften aus. Es entstand dabei ein Kategoriensystem mit 19 Inhaltskategorien (z. B. Tiere, Berufe, Spielzeug, Aktionen …), anhand derer eine möglichst gleichmäßige inhaltliche Verteilung der Aufgaben und eine breite Abdeckung des Wortschatzes erzielt wurden. Die dritte Fassung enthielt je Testfassung 204 Aufgaben.

In Fassung 4 wurden schließlich Boden- und Deckeneffekte durch Hinzunahme sehr leichter und sehr schwerer Aufgaben weitgehend reduziert. Außerdem erfolgte eine komplette Überarbeitung des Bildmaterials und eine bessere Ausbalancierung hinsichtlich der Inhaltskategorien. So wurde der Anteil an Aufgaben in den Kategorien *Körperteile*, *Kleidung und Zubehör*, *Obst und Gemüse*, *Musikinstrumente und Spielzeug* und *Erholung* erhöht, der Anteil der Kategorien *Aktionen* und *Adjektive* hingegen reduziert. Beide Maßnahmen dienten dem Zweck, das Testverfahren stärker für die Anwendung bei jüngeren Kindern zu optimieren.

Bei der Adaptation ins Deutsche wurde die Übersetzung jedes einzelnen Wortes von Personen mit jeweils deutscher oder englischer Muttersprache gemeinsam vorgenommen. Bei jenen Zielwörtern, bei denen keine stimmige Übertragung möglich war, wurde eines der Distraktorbilder als Target verwendet, die Wortart geändert (z. B. coniferous → Konifere) oder auf einen unterschiedlichen Aspekt des Targetbildes fokussiert. Das Bildmaterial wurde also komplett beibehalten.

Ein besonderer Fokus lag sowohl in der amerikanischen als auch in der deutschen Version in der gerechten Ausgestaltung der Stimuli. So sind Menschen beiderlei Geschlechts und verschiedener Ethnien gleichmäßig auf alle Berufsgruppen und Tätigkeiten aufgeteilt. Die Testfairness wurde mittels Item Response Theory (IRT)- und Differential Item Functioning (DIF)-Analysen untersucht. Mit Letzteren lässt sich überprüfen, ob einzelne Aufgaben für bestimmte Bevölkerungsgruppen unverhältnismäßig leicht oder schwer sind, wodurch die Fairness des Verfahrens negativ beeinflusst werden könnte. Dabei geht es nicht darum, ob eine Bevölkerungsgruppe bei einem Wort im Schnitt besser oder schlechter abschneidet als eine andere Gruppe, sondern ob die Lösungswahrscheinlichkeit für dieses Wort bei einer Bevölkerungsgruppe niedriger oder höher liegt, als auf der Basis des durchschnittlichen Fähigkeitsniveaus dieser Gruppe erwartet werden würde. Ist dies der Fall, so kann man schließen, dass die Lösung außer der Fähigkeit durch weitere Merkmale (z. B. dem Sprachhintergrund) beeinflusst wird (vgl. Swaminathan & Rogers 1990). In der US-Fassung wurde diese Analyse für die Variablen Geschlecht, ethnische Herkunft, Sozialstatus und Wohnort innerhalb der USA durchgeführt, in der deutschsprachigen Adaptation für die Variablen Geschlecht und Sprachhintergrund. Damit sollte unter anderem ein möglichst hoher Grad an Kulturunabhängigkeit erreicht werden.

Da die maßgebliche Motivation für die Erstellung der deutschsprachigen Fassung die Vergleichbarkeit der Ergebnisse zwischen der englischen und der deutschen Sprache war, wurden die einzelnen Wörter mit möglichst gleichfrequenten Wörtern der deutschen Sprache übersetzt. Dabei kam ein relatives Maß der Worthäufigkeit zum Einsatz, bei dem die Frequenz des häufigsten Wortes der jeweiligen Sprache durch die Auftretenshäufigkeit des Zielwortes geteilt und der Quotient logarithmiert wird. Zwar wird auf diese Weise

eine gemeinsame, sprachübergreifende Metrik etabliert, jedoch gibt es dennoch prinzipielle Unterschiede zwischen der deutschen und der englischen Sprache, die beispielsweise in den Wurzeln beider Sprachen begründet sind. Obwohl beide Sprachen zu den germanischen Sprachen zählen und erhebliche Verwandtschaft aufweisen, kommen in der amerikanischen Version des PPVT-4 einige Wörter vor, für die das Deutsch keine passende Übersetzung mit ähnlicher Wortschwierigkeit bereithält. Ein Beispiel ist das englische Wort "cairn", das aus dem Gälischen stammt und pyramidenförmig aufgeschichtete Steinhügel des Neolithikums bezeichnet, die nur auf den britischen Inseln vorkommen. Man könnte den Begriff zwar schlicht mit "Steinhügel" übersetzen, würde damit jedoch eine deutlich andere Schwierigkeit erzielen, da "cairn" fast ausschließlich fachsprachliche Anwendung findet.

Weiterhin verfügt die englische Sprache über ein größeres Vokabular und damit einhergehend über mehr Synonyme unterschiedlicher Schwierigkeit. So führt beispielsweise das Oxford English Dictionary aktuell über 600.000 Schlagwörter auf, während die Duden-Redaktion den deutschen Wortschatz nur auf 300.000 bis 400.000 beziffert (Kunkel-Razum 2000). Die Ursache hierfür liegt mutmaßlich in der Eroberung der britischen Inseln durch die Normannen 1066, wodurch das heutige Englisch sowohl einen großen germanischen als auch romanischen Wortschatz hat. Wörter mit romanischer Wurzel wurden zunächst jedoch eher im höfischen Kontext gebraucht, wohingegen Wörter mit germanischer Wurzel in die Alltagssprache niedrigerer gesellschaftlicher Klassen abgedrängt wurden. Auch heute noch werden Wörter mit germanischer Wurzel tendenziell eher in alltäglichen oder profanen Situationen gebraucht, während ihre romanischen Entsprechungen einen höheren Grad an Elaboriertheit und eine niedrigere Frequenz aufweisen. Im PPVT-4 betrifft dies beispielsweise das Wort "beverage" (= Getränk), welches mit dem französischen Verb boire (= trinken) verwandt ist (vgl. auch franz. "beuverie" = Zecherei). Das germanischstämmige Synonym dazu lautet "drink". Für den deutschsprachigen Begriff "Getränk" stehen dagegen kaum vergleichbare Alternativen zur Verfügung. Dadurch ist die Schwierigkeit jedes einzelnen Wortes im Englischen höher als diejenige des deutschen Begriffs "Getränk", da jedes Synonym jeweils weniger häufig verwendet wird als sein deutsches Pendant. Deutschsprachige Kinder lösen aufgrund der geringeren lexikalischen Diversität der Sprache deshalb bei der deutschen Adaptation des PPVT-4 im Schnitt mehr Aufgaben richtig als gleichaltrige englischsprachige Kinder in der US-Version. Die deutschsprachige Testversion differenziert in den höheren Altersstufen und im hohen Leistungsbereich also etwas weniger gut als die amerikanische. Die Adaption des PPVT-IV fand in Deutschland statt, sodass die Testgüte für die Schweiz nicht systematisch untersucht wurde, jedoch stehen mittlerweile weiterführende Erfahrungen zur Verfügung (siehe Juska-Bacher & Röthlisberger 2021). Mit leichten Anpassungen (z.B. Item 65: Schornstein –

Kamin, Item 70: Umschlag – Kuvert) lässt sich das Testverfahren auch erfolgreich in der Deutschschweiz einsetzen.

3.3 Normierung und Reliabilität

Die deutsche Übersetzung des PPVT-4 wurde Anfang 2013 an einer Stichprobe von N = 389 Kindern der Klassenstufe 3 bis 10 aus Grund-, Hauptschulen und Gymnasien überprüft. Die Darbietung der vier Auswahlbilder pro Aufgabe erfolgte ohne Abbruchkriterien mittels eines dafür erstellten Computerprogramms, welches auch die Antwortzeiten protokollierte. Das zugehörige Wort, das richtig zugeordnet werden musste, wurde auditiv über Kopfhörer vorgespielt. Für die Neunormierung des Tests wurden die Items gemäß der im Deutschen erzielten Schwierigkeiten und – bei gleicher Schwierigkeit – der Reaktionszeiten aufsteigend gereiht.

Diese endgültige Testversion wurde schließlich zur Erhebung der deutschen Normierungsdaten verwendet. Hierfür führten 42 Testleiter_innen das Verfahren gemäß der standardisierten Testdurchführung in 20 Kindergärten und 47 Schulen in ganz Deutschland durch. Insgesamt nahmen 4880 Kinder und Jugendliche im Alter zwischen 2;7 Jahren und 18;0 Jahren an der Untersuchung teil. Die Stichprobe wurde schließlich für jede Klassenstufe sowie für Kinder aus dem Kindergarten hinsichtlich Geschlecht, Schulart und Anteil an Kindern mit Migrationshintergrund stratifiziert. Bei Überrepräsentation von Merkmalen infolge der geclusterten Datenerhebung erfolgte eine zufällige Ziehung von Fällen bis die Zielquote erreicht wurde. In der repräsentativen Normstichprobe verblieben schließlich N = 3555 Kinder und Jugendliche aus den Bundesländern Baden-Württemberg, Bayern, Brandenburg, Hessen, Niedersachsen, Nordrhein-Westfalen und Rheinland-Pfalz. Der Anteil an Kindern und Jugendlichen mit Migrationshintergrund im Sinne mindestens eines im Ausland geborenen Elternteils betrug 29.3 %, was gemäß des Zensus als repräsentativ angesehen werden kann. Die endgültige Testversion liefert Normen für Kinder zwischen 3;0 und 16;11 Jahren. Die einzelnen Altersgruppen umfassen dabei zwischen 102 und 431 Kinder pro Jahrgang, wobei die Altersnormen mit einem kontinuierlichen Normierungsverfahren modelliert wurden (A. Lenhard et al. 2018).

Für die Beurteilung der Testgüte wurden die Rohdaten mittels eines 1PL-IRT-Modells skaliert. Eine Analyse der tatsächlich bearbeiteten Aufgaben (Aufgaben unterhalb des Boden- und oberhalb des Deckensets wurden dabei als Missing behandelt) ergab sehr hohe Reliabilitätskennwerte mit einer EAP- und RLE-Reliabilität von r = .965. Auch die Odd-Even-Split-Half-Reliabilitäten lagen aufgeteilt nach Altersgruppe oder Klassenstufe durchgehend bei mindestens r_{tt} = .92, d. h. im sehr guten Bereich (siehe A. Lenhard et al. 2015: 76). Dies galt ebenso für die Retestreliabilität, die nach 6 bis 12 Monaten r_{tt} = .91 betrug.

3.4 Altersentwicklung und Heterogenität des rezeptiven Wortschatzes

Den absoluten Umfang des Wortschatzes einer Person zu bestimmen, ist messtechnisch nur sehr schwierig möglich, insbesondere da es nicht nur auf die Kenntnis eines Wortes, sondern auch auf die Tiefe des Wissens ankommt. Die Wortschatztiefe spiegelt sich beispielsweise nicht nur in der Kenntnis, sondern auch in der Organisation des Wortschatzes, in der Fähigkeit zum rezeptiven und produktiven Einsatz, in der Schnelligkeit beim Zugriff auf die Wortbedeutung usw. wider (Schmitt 2014). Diese Aspekte sind unscharf und überlappend und deswegen schwer voneinander zu trennen. Die Verfügbarkeit eines Wortes im mentalen Lexikon ist nicht dichotom (verfügbar versus nicht verfügbar), sondern graduell, beginnend bei der Wiedererkennung der phonetischen Form oder visuellen Gestalt des geschriebenen Wortes ohne Kenntnis der Bedeutung bis hin zu einem gereiften Verständnis und tiefergehenden Wissen, möglicherweise sogar der etymologischen Wurzeln. In den meisten anwendungsbezogenen Fällen wird das rezeptive Wiedererkennen von Wörtern im Sinne der Zuordnung zu einem Bedeutungsgehalt als ein hinreichend guter Indikator für die Größe und Tiefe des Wortschatzes angesehen. So korrelierte beispielsweise in der Untersuchung von Vermeer (2001) zum Wortschatz bei Erst- und Zweitsprache in einer Gruppe niederländischer Kinder der rezeptive und produktive Wortschatz zu $r = .80$, auch wenn in absoluten Zahlen der rezeptive Wortschatz – erhoben mit einem Instrument, dessen Aufbau dem PPVT-4 entsprach – erheblich größer war als der produktiv eingesetzte. Auch beim PPVT-4 spezifizieren die Rohwerte die absolute Größe des Wortschatzes zwar nicht direkt, dienen aber als Indikator für die relative Größe des Wortschatzes im Vergleich zu Personen des gleichen Alters. Um diesen Vergleich zu ermöglichen, verfügt der PPVT-4 analog zu anderen psychometrischen Untersuchungsinstrumenten über Normwerte, welche die Ausprägung der latenten Wortschatzgröße und die relative Position der Person im Vergleich zur Normierungsstichprobe wiedergeben (siehe Kap. 3.3). Der Verlauf dieser Normwerte ist in Form einiger ausgewählter Perzentilbänder in Abbildung 2 dargestellt.

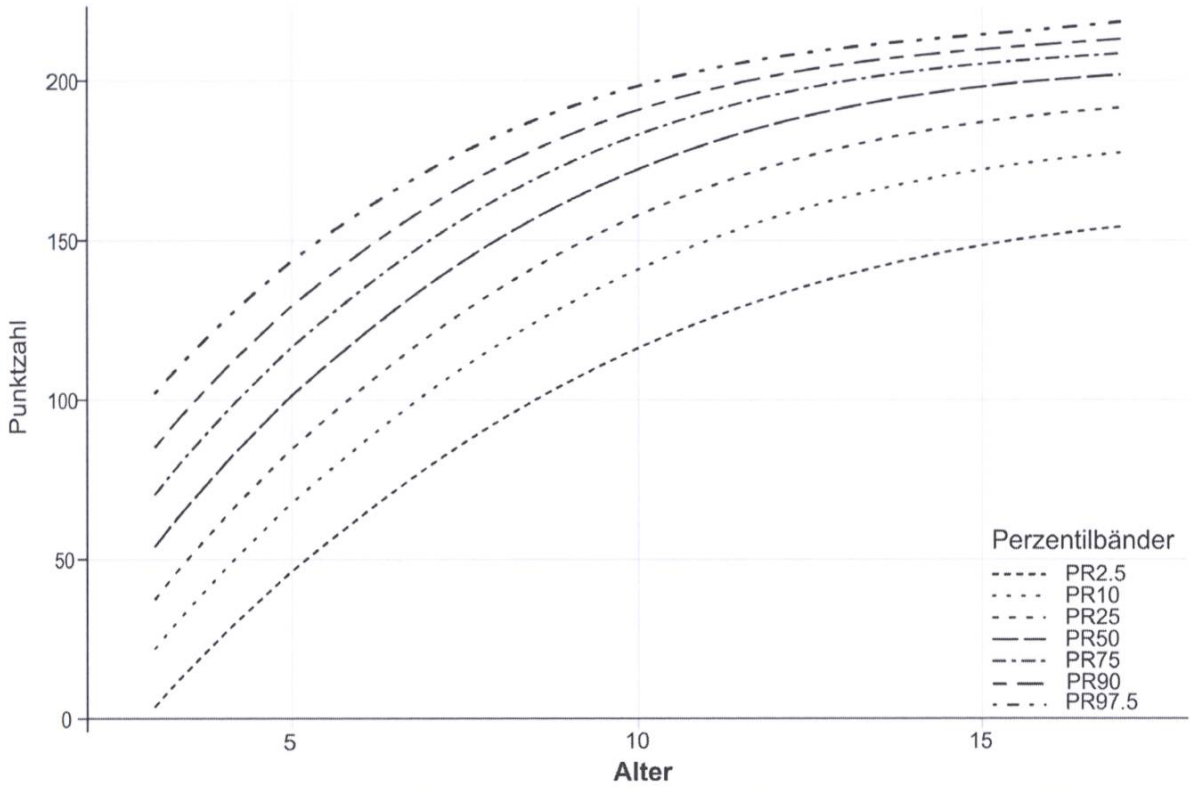

Abbildung 2: Perzentilbänder der Normierungsstichprobe im Altersverlauf von 3 bis 17 Jahren

Der Verlauf der Perzentilbänder zeigt erwartungsgemäß einen starken Anstieg des mittleren Wortschatzes, der besonders im Altersbereich von 3 bis 10 Jahren steil verläuft (siehe auch Tab. 1, Haupteffekte Alter). Die anschließende Abflachung spiegelt nicht ausschließlich eine geringere Entwicklungsgeschwindigkeit ab der frühen Pubertät wider, sondern ist auch der Verteilung der Aufgabenschwierigkeiten im Testverfahren geschuldet. Man beachte allerdings, dass es sich bei der Darstellung nicht um längsschnittliche Verläufe, sondern um die querschnittliche Modellierung der Normierungsdaten handelt. Nichtsdestotrotz kann im erfassten Altersbereich unabhängig von der Fähigkeitsausprägung von einem kontinuierlichen Zuwachs des Wortschatzes ausgegangen werden. Der Alterseffekt ist sehr stark ausgeprägt, F (3, 3351) = 2916.2, $p < .001$ und erklärt mit $\eta^2_{partial} = .72$ einen erheblichen Teil der Varianz der Rohwerte.

		Punktzahl			
Prädiktor	***Schätzwert***	***CI***	***t***	***SE***	***p***
Haupteffekte					
(Intercept)	173.21	170.82 – 175.60	1.22	141.96	**<0.001**
Alter (linear)	1707.03	1586.61 – 1827.45	61.42	27.79	**<0.001**
Alter (quadratisch)	-559.10	-596.89 – -521.31	19.27	-29.01	**<0.001**
Alter (kubisch)	114.42	77.05 – 151.78	19.06	6.00	**<0.001**
MI einfach	-12.91	-22.20 – -3.62	4.74	-2.73	**0.006**
MI beide	-35.68	-44.35 – -27.00	4.42	-8.07	**<0.001**
Geschlecht	-3.07	-7.56 – 1.41	2.29	-1.34	0.179
Interaktionseffekte					
MI einfach x Alter	0.67	0.02 – 1.33	0.33	2.02	**0.043**
MI beide x Alter	0.55	-0.04 – 1.14	0.30	1.83	0.067
Geschlecht x Alter	-0.01	-0.42 – 0.40	0.21	-0.06	0.954
MI einfach x Geschlecht	-1.53	-5.49 – 2.43	2,021	-0,76	0.449
MI beide x Geschlecht	5.46	1.90 – 9.02	1,815	3,01	**0.003**
Anzahl Fälle	3363				
R^2	0.736				

Tabelle 1: Ergebnisse einer polynomialen Regression der Haupteffekte Alter, Geschlecht und Migrationshintergrund (MI) auf die erzielte Punktzahl[1]

Ins Auge fällt zudem die sehr breite Verteilung der Ergebnisse innerhalb jeder Altersstufe. So entspricht ein Rohwert von 49 im Alter von 5 Jahren einem Prozentrang von 2.5 und ein Rohwert von 145 einem Prozentrang von 97.5. Die letztgenannte Leistung erreichen die leistungsschwächsten 2.5 % erst im Alter von 15 Jahren. Anders ausgedrückt besitzen die sehr leistungsstarken fünfjährigen Kinder einen Entwicklungsvorsprung von etwa 10 Jahren im Vergleich zu den sehr leistungsschwachen!

[1] MI bezeichnet den Migrationshintergrund der Eltern (einfach = ein Elternteil ist migriert; beide = beide Elternteile sind migriert) mit Personen ohne Migrationshintergrund der Eltern als Referenzgruppe. Signifikante Einflussfaktoren sind hervorgehoben. Die Schätzwerte geben den mittleren Einfluss der unabhängigen Variablen auf den Gesamtrohwert an. Positive Schätzwerte erhöhen folglich im Schnitt das erzielte Ergebnis, negative Werte senken es. Bei Kodierung des Geschlechts bildeten die Jungen die Referenzkategorie.

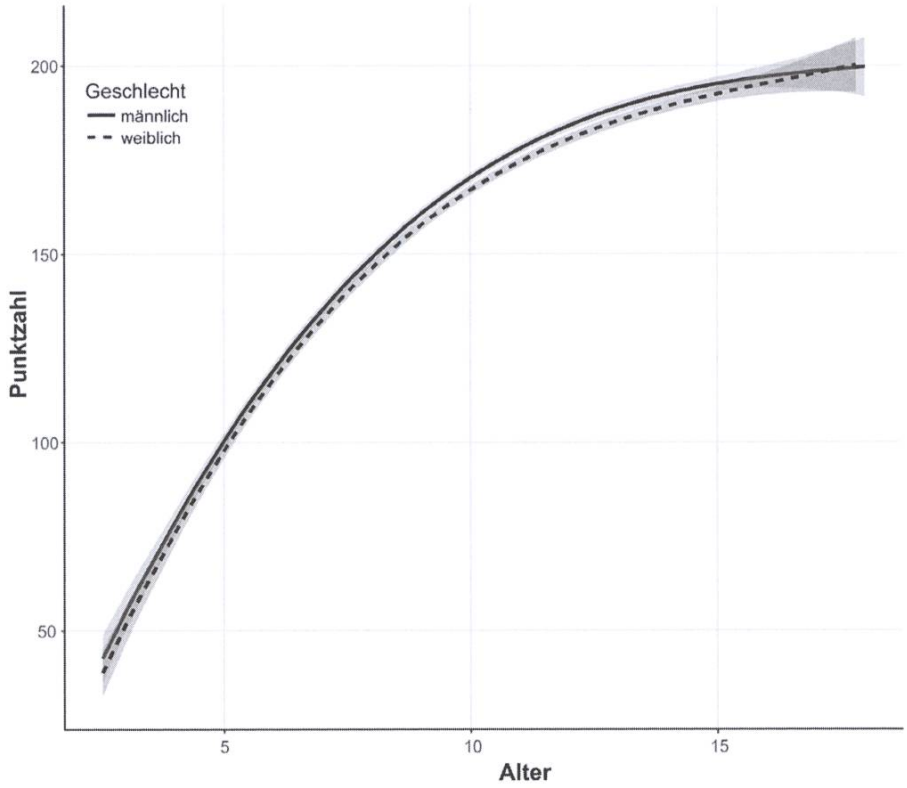

Abbildung 3: Mittlere Leistung im Altersverlauf nach Geschlecht[2]

3.5 Effekte von Geschlecht, Migrationshintergrund und Alter

Jungen und Mädchen der Normierungsstichprobe zeigten im Schnitt vergleichbare Leistungen (Abbildung 3, Haupteffekt Geschlecht, Tab. 1), mit einem minimalen Nachteil der Mädchen. Dieser Unterschied war allerdings so gering, dass er keine statistische Bedeutsamkeit erreichte und lediglich 1.9 ‰ der Varianz aufklärt. Auch fanden sich keine relevanten Interaktionen von Alter und Geschlecht, sodass der Unterschied über die Altersspanne im wesentlich konstant blieb. Diese Befunde galten sowohl für Menschen ohne Migrationshintergrund als auch für Personen mit einem migrierten Elternteil. Bei Personen, bei denen beide Eltern eingewandert waren, wechselwirkte das Geschlecht hingegen mit den Sprachfertigkeiten dahingehend, dass Jungen einen niedrigeren Wortschatz aufwiesen (Tab. 1, Interaktionseffekt MI beide x Geschlecht) und dieser Nachteil vor allem im leistungsschwachen Bereich stark ausgeprägt war. Der Unterschied zwischen Mädchen und Jungen betrug in dieser Gruppe im Schnitt 5.45 Rohpunkte. Mädchen wiesen also, anders als auf der Basis von Geschlechtsstereotypen zu erwarten gewesen wäre, nicht generell einen größeren Wortschatz auf – ein Befund, der im Rahmen der Intelligenzforschung bereits als gut gesichert gilt (Toivainen et al. 2017). Allem

2 Die Linien geben die mittlere Leistung von Jungen (durchgezogene Linie) und Mädchen in Abhängigkeit vom Alter wieder, die grauen Bänder stellen die 95%-Konfidenzinterballe dar.

Anschein nach existieren jedoch sehr wohl einzelne Gruppen unter den Jungen, die besonders benachteiligt sind.

Eine Migrationserfahrung der Eltern hat einen sehr starken Einfluss auf die Wortschatzleistung, $F(2, 3351) = 288.8$, $p < .001$, $\eta^2_{partial} = .15$. Personen, bei denen ein Elternteil eingewandert war und die i. d. R. bilingual aufwuchsen, erzielten eine im Schnitt 12.9 niedrigere Punktzahl. Besonders stark fiel der Unterschied bei Personen aus, bei denen beide Eltern eingewandert waren. Sie lagen im Schnitt 35.7 Punkte hinter den Personen ohne Migrationshintergrund – ein Unterschied, der je nach Altersgruppe zwischen 1.5 und 2.0 Standardabweichungen beträgt und als sehr groß eingestuft werden kann. In beiden Personengruppen reduzierte sich der Unterschied zu Personen ohne Migrationshintergrund in den höheren Altersgruppen, wobei diese Interaktion nur bei den Personen mit einem migrierten Elternteil Signifikanz erreicht. Insbesondere die Altersverläufe in dieser Gruppe zeigen eine sehr spannende Form: Während in jüngeren Kohorten die Leistung stärker den Kindern mit zwei migrierten Elternteilen gleicht, reduziert sich der Leistungsnachteil zum Schuleintritt und bleibt dann lange konstant. In der Pubertät schließlich nähern sich Personen mit einem migrierten Elternteil immer stärker der Kohorte ohne Migrationserfahrung an und überflügeln diese am Ende der Jugend deskriptiv sogar leicht (Abbildung 4). Die Gruppe an Kindern mit zwei migrierten Elternteilen holt dagegen nur bis zum Alter von 10 Jahren leicht auf, weist danach aber einen konstanten Abstand zu Personen ohne Migrationshintergrund auf, sodass sich in der Interaktion von Migrationshintergrund und Alter (siehe Tabelle 1) nur ein Trend zeigt.

Die Verteilung der Rohdaten der Person mit zwei migrierten Elternteilen zeigt innerhalb der einzelnen Altersgruppen eine größere Schiefe als in den anderen Gruppen. Ähnlich wie bei den Geschlechtseffekten fallen leistungsschwache Personen überproportional zurück. Der Abstand der Perzentile der Personen mit zwei migrierten Elternteilen zu den entsprechenden Referenzperzentilen der monolingual deutschsprachigen Personen nimmt also mit sinkender Fähigkeit immer weiter zu.

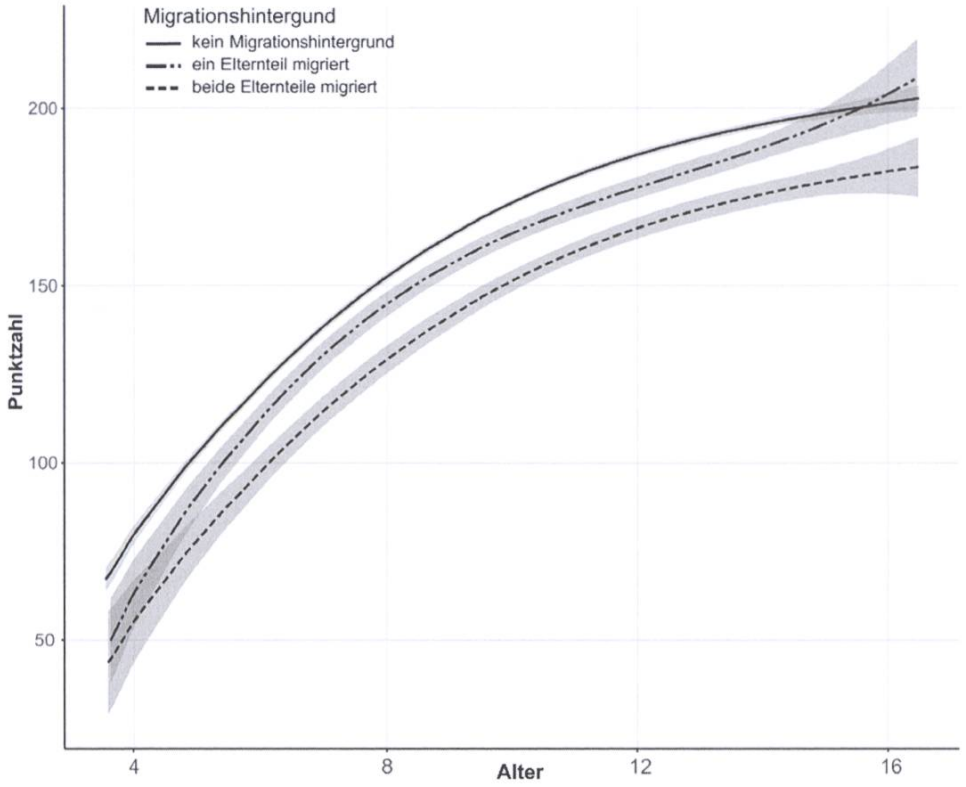

Abbildung 4: Der Einfluss der Familiensprache auf die im Mittel erzielten Rohpunkte[3]

4. Abschluss

Abschließend möchten wir auf die Ziele und Grenzen einer Wortschatzdiagnostik mit dem PPVT-IV eingehen. Eine offensichtliche Grenze liegt darin, dass natürlich keineswegs überprüft wird, welche Wörter im Wortschatz fehlen, sodass das Verfahren nicht für die Feindiagnostik zur Planung von Wortschatzeinheiten im Rahmen von formellem Deutschunterricht geeignet ist. Hierfür müssten in der Tat jene Wörter überprüft werden, deren Beherrschung erreicht werden soll. Stattdessen zielt das Verfahren darauf ab, über die Normierung die relative Position einer Person innerhalb der Altersgruppe zu bestimmen. Es dient damit als Diagnoseinstrument zur Ermittlung des Sprachstandes, als international anerkannter psychometrischer Test für Forschungszwecke, also beispielsweise zum Vergleich des Wortschatzerwerbs in verschiedenen Sprachen, oder auch um die Ursache schulischen Scheiterns genauer zu untersuchen. So lässt sich beispielsweise mit Hilfe des PPVT-4 differenzialdiagnostisch überprüfen, ob schlechte Leistungen im Lesen und Schreiben bei einem Kind, das nicht oder nicht ausschließlich Deutsch als Familiensprache spricht, auf die Schriftsprache

[3] Die Linien geben die mittlere Leistung von Personen ohne Migrationshintergrund (Referenzgruppe, Code 1), Personen mit einem migrierten Elternteil und Personen mit zwei migrierten Elternteilen in Abhängigkeit vom Alter wieder, die grauen Bänder stellen die 95%-Konfidenzinterballe dar.

beschränkt sind, oder ob diese Probleme eventuell durch allgemeine Sprachrückstände bedingt sind.

LITERATUR

Beck, I. L., Perfetti, C. A. & McKeown, M. G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology*, *74(4)*, 506-521.

Brugger, L. & Juska-Bacher, B. (2021). Assessing primary grade children's lexical inferencing strategies while reading – A review. *Bulletin suisse de linguistique appliquée, 113,* 213-232.

Bynner, J. & Parsons, S. (2005). *New light on literacy and numeracy*. London: National Research and Development Centre for adult literacy and numeracy.

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

Cromley, J. G. & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99(2)*, 311-325.

Eckes, T. & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, *23(3)*, 290-325.

Hamed, O. & Zesch, T. (2018). The automatic generation of nonwords for lexical recognition tests. In Z. Vetulani, J. Mariani, & M. Kubis (eds.), *Lecture notes in computer science. Human language technology. Challenges for computer science and linguistics* (Vol. 10930, pp. 321-331). Cham: Springer.

Hattie, J. (2010). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.

Heller, K. A. & Perleth, C. (2000). *KFT 4-12+ R kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision*. Göttingen: Beltz.

Jäger, A.O., Holling, H., Preckel, F., Schulze, R., Vock, M., Süß, H.-M. & Beauducel, A. (2006). *BIS-HB. Berliner Intelligenzstruktur-Test für Jugendliche*. Göttingen: Hogrefe.

Juska-Bacher, B. & Röthlisberger, M. (2021). Das Konstrukt Wortschatz: Dimension(en) Umfang und Tiefe? Empirische Ergebnisse aus der Unterstufe. *Bulletin suisse de linguistique appliquée, 113,* 49-68.

Khodadady, E. (2014). Construct validity of c-tests: A factorial approach. *Journal of Language Teaching and Research, 5(6)*, 1353-1362.

Kunkel-Razum, K. (2000). *Wie viele Wörter hat die deutsche Sprache?* Goethe-Institut. Verfügbar unter: https://www.goethe.de/ins/be/de/kul/prj/ssk/21784921.html

Lemhöfer, K. & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods, 44(2)*, 325-343.

Lenhard, A. (2020). *Das aufgezeichnete Video-Seminar zur WISC-V* [Videoseminar]. Frankfurt a. M.: Pearson.

Lenhard, A., Lenhard, W., Segerer, R. & Suggate, S. (2015). *Peabody Picture Vocabulary Test - Revision 4* (PPVT-4), deutsche Version. Frankfurt a. M.: Pearson Assessment.

Lenhard, A., Lenhard, W., Suggate, S. & Segerer, R. (2018). A continuous solution to the norming problem. *Assessment*, *25(1)*, 112-125.

Lenhart, J., Suggate, S. & Lenhard, W. (im Druck). Shared-reading onset and emergent literacy development. *Early Education and Development*.

Pfost, M., Dörfler, T. & Artelt, C. (2010). Der Zusammenhang zwischen außerschulischem Lesen und Lesekompetenz. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie, 42(3)*, 167-176.

Pfost, M., Dörfler, T. & Artelt, C. (2013). Students' extracurricular reading behavior and the development of vocabulary and reading comprehension. *Learning and Individual Differences, 26*, 89-102.

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F. & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence, 53*, 118-137.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning, 64(4)*, 913-951.

Schneider, W. J. & McGrew, K. S. (2018). The Catell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan & McDonough, E. M. (eds.), *Contemporary intellectual assessment* (pp. 73-163). New York: the Guilford Press.

Sims, V. M. (1929). The reliability and validity of four types of vocabulary tests. *The Journal of Educational Research, 20(2)*, 91-96.

Suggate, S. P., Reese, E., Lenhard, W. & Schneider, W. (2014). The relative contributions of vocabulary, decoding, and phonemic awareness to word reading in English versus German. *Reading and Writing, 27*, 1395-1412.

Suggate, S., Lenhard, W., Neudecker, E. & Schneider, W. (2013). Incidental vocabulary acquisition from stories: Second and fourth graders learn more from listening than reading. *First Language, 33(6)*, 551-571.

Suggate, S., Lenhart, J., Vaahtorana, E. & Lenhard, W. (2021). Interactive elaborative story-telling fosters vocabulary in preschoolers compared to repeated-reading and phonemic awareness interventions. *Cognitive Development*, *57*, Artikel 100996.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27(4)*, 361-370.

Toivainen, T., Papageorgiou, K. A., Tosto, M. G. & Kovas, Y. (2017). Sex differences in non-verbal and verbal abilities in childhood and adolescence. *Intelligence, 64*, 81-88.

Trautwein, J. & Schroeder, S. (2019). WOR-TE: Ein Ja / Nein-Wortschatztest für Kinder verschiedener Altersgruppen. *Diagnostica, 65(1)*, 37-48.

Vaahtoranta, E., Lenhart, J., Suggate, S. & Lenhard, W. (2019). Interactive elaborative storytelling: Engaging children as storytellers to foster vocabulary. *Frontiers in Psychology, 10,* 1534.

Verhoeven, L., van Leeuwe, J. & Vermeer, A. (2011). Vocabulary growth and reading development across the elementary school years. *Scientific Studies of Reading, 15(1)*, 8-25.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*, 217-234.

Wechsler, D. (2008). *Wechsler intelligence scale for adults (4th ed.)*. Bloomington, MN: NCS Pearson.

Wechsler, D. (2014). *Wechsler intelligence scale for children (5th ed.)*. Bloomington, MN: NCS Pearson.