Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

MEZ
Mehrsprachigkeitsentwicklung
im Zeitverlauf

MEZ Arbeitspapiere

Thorsten Klinger, Birger Schnoor

Prüfung der Messinvarianz über die Zeit: Das Konstrukt Schreibfähigkeit im Deutschen

MEZ Arbeitspapier Nr. 6
Hamburg, Februar 2020

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Arbeitspapiere des Forschungsprojekts
Mehrsprachigkeitsentwicklung im Zeitverlauf – MEZ
an der Universität Hamburg
In der Reihe erscheinen Beiträge zu Themen, die den Arbeitsprozess des MEZ-Projekts betreffen. Die Beiträge erscheinen während der Projektlaufzeit und berichten vorläufige Ergebnisse zu den theoretischen und empirischen Fragestellungen des Projekts sowie Lösungen zum technischen Umgang mit den Projektdaten.

Principal Investigators MEZ-Projekt:
Ingrid Gogolin, Universität Hamburg (Projektleitung)
Christoph Gabriel, Johannes-Gutenberg-Universität Mainz
Michel Knigge, Universität Potsdam
Marion Krause, Universität Hamburg
Peter Siemund, Universität Hamburg

Bezug:
www.mez.uni-hamburg.de

Kontakt:
Mehrsprachigkeitsentwicklung im Zeitverlauf – MEZ
Universität Hamburg
Institut für Interkulturelle und International Vergleichende Erziehungswissenschaft
Postanschrift: Von-Melle-Park 8, 20146 Hamburg
Besucheranschrift: Alsterterrasse 1, 5. Stock, 20354 Hamburg
E-Mail: mez@uni-hamburg.de
Tel.: +49 40 42838-3950

Bitte zitieren Sie dieses Arbeitspapier wie folgt:
Klinger, Thorsten; Schnoor, Birger (2020): Prüfung der Messinvarianz über die Zeit:
Das Konstrukt Schreibfähigkeit im Deutschen. MEZ Arbeitspapier Nr. 6. Hamburg (Universität Hamburg): www.mez.uni-hamburg.de

Thorsten Klinger, Birger Schnoor

Prüfung der Messinvarianz über die Zeit: Das Konstrukt Schreibfähigkeit im Deutschen

MEZ Arbeitspapier Nr. 6

Zusammenfassung:

Im Rahmen des vom BMBF geförderten Projekts „Mehrsprachigkeitsentwicklung im Zeitverlauf (MEZ)“ wurde die sprachliche Entwicklung von rund 2000 Sekundarstufenschüler(innen) der Jahrgänge sieben und neun über einen Zeitraum von drei Jahren untersucht. Die Erhebungen an der Gesamtstichprobe mit vier Erhebungswellen beziehen sich auf die rezeptiven und produktiven schriftsprachlichen Fähigkeiten in den Sprachen Deutsch, den Herkunftssprachen Russisch bzw. Türkisch sowie in den Schulfremdsprachen Englisch und gegebenenfalls Französisch bzw. Russisch. Das vorliegende Arbeitspapier bezieht sich auf die Messung produktiver Schreibfähigkeiten im Deutschen. Es wird geprüft, mit welchem Grad an Präzision das longitudinale Messmodell die Entwicklung der Schreibfähigkeiten im Deutschen misst. Im Vordergrund steht die Frage, ob gemessene Veränderungen in den Schreibfähigkeiten über die Zeit verlässlich auf Veränderungen in der Schreibkompetenz zurückzuführen sind oder auf andere Varianzquellen.

Schlagworte:

Messung von Schreibfähigkeiten im Deutschen – konfirmatorische Faktorenanalyse – Messinvarianz über die Zeit

Abstract:

The project "Multilingual Development: A Longitudinal Perspective (MEZ)", funded by the German Federal Ministry for Education and Research (BMBF), examined the language development of about 2000 secondary school students (grades seven and nine) over a three years period. The surveys on the overall sample (four survey waves) cover receptive and productive written-language skills in German, the heritage languages Russian and Turkish, in the foreign languages English and where applicable French and Russian. This working paper refers to the measurement of productive writing skills in German. It examines the measurement model's degree of precision in assessing the development of writing skills in German over time. The key question is whether measured changes in writing skills over time can be reliably attributed to changes in writing competence or to other sources of variance.

Keywords:

assessment of writing skills in German – confirmatory factor analysis – measurement invariance across time

Inhaltsverzeichnis

1. Einleitung

Das Projekt „Mehrsprachigkeitsentwicklung im Zeitverlauf (MEZ)“[1] untersucht die sprachliche Entwicklung von lebensweltlich mehrsprachigen und monolingual deutschsprachigen Jugendlichen. Die Längsschnittstudie zielt auf die Beantwortung der Fragen, welche sprachlichen, personalen und kontextuellen Bedingungen die Aneignung von Mehrsprachigkeit positiv oder negativ beeinflussen, wie sich diese Bedingungen über die Zeit verändern, in welchen Wechselbeziehungen sie zu einander stehen sowie welche Zusammenhänge dabei mit der schulischen Entwicklung und der Vorbereitung auf einen beruflichen Werdegang bestehen (Gogolin et al. 2017). An insgesamt vier Messzeitpunkten wurden an allgemeinbildenden Schulen in acht deutschen Bundesländern über 2000 Schüler(innen) der Jahrgänge sieben und neun über einen Zeitraum von drei Jahren verfolgt (Brandt et al. 2017). Die Untersuchung der sprachlichen Entwicklung an der Gesamtstichprobe bezieht sich auf die rezeptiven und produktiven schriftsprachlichen Fähigkeiten in den Sprachen Deutsch, den Herkunftssprachen Russisch bzw. Türkisch sowie in den Schulfremdsprachen Englisch und gegebenenfalls Französisch bzw. Russisch (Gogolin et al. 2017).

Die MEZ-Studie folgt einem sprachtheoretisch orientierten Verständnis von Sprachkompetenz, deren Teilbereiche als eigenständige Fähigkeiten aufgefasst werden und deren Beziehungen zueinander theoretisch und empirisch zu klären sind (Klinger et al. 2019). Dabei stehen vor allem schriftsprachliche Fähigkeiten im Vordergrund des Interesses, denen nicht nur in bildungsbezogenen Kontexten eine grundlegende Bedeutung zukommt.

Literalität bezieht sich sowohl auf die rezeptive Fähigkeit zu lesen als auch auf die produktive Fähigkeit zu schreiben. Während in bildungswissenschaftlichen Large-Scale-Untersuchungen rezeptive Sprachfähigkeiten wie Lese- und Hörverstehen (auch als Annäherung an sprachliche Fähigkeiten insgesamt) dominieren (z. B. Wendt et al. 2016; Hußmann et al. 2017; Reiss et al. 2019), wird die produktive Schreibkompetenz in großangelegten Bildungsstudien nur selten berücksichtigt (z. B. DESI-Konsortium 2008). MEZ bezieht dagegen ausdrücklich neben Lesefähigkeiten auch Fähigkeiten zur Produktion schriftsprachlicher Texte ein (vgl. Usanova und Klinger 2020), was eine Einschätzung bildungsrelevanter Kompetenzen über die üblicherweise gemessenen rezeptiven sprachlichen Fähigkeiten (Lesefähigkeit) hinaus ermöglicht.

Das vorliegende Arbeitspapier bezieht sich auf die Messung produktiver Schreibfähigkeiten in der MEZ-Studie. Zu diesem Zweck wird zunächst ein kurzer Überblick über das dahinterstehende Sprachkonstrukt und die darauf bezogenen Erhebungs- und Auswertungsinstrumente gegeben (ausführlicher dazu: Usanova und Klinger 2020).

Im Mittelpunkt des Arbeitspapiers steht das daraus resultierende longitudinale Messmodell für die Schreibfähigkeiten, das sich hier zunächst auf die Fähigkeiten im Deutschen beschränkt. Dabei werden die Messungen aller vier Messzeitpunkte einbezogen. In unserem Beitrag wird geprüft, mit welchem Grad an Präzision das Messmodell die Entwicklung der Schreibfähigkeiten im Deutschen misst. Im Vordergrund steht die Frage, ob gemessene Veränderungen in den Schreibfähigkeiten über die Zeit verlässlich auf Veränderungen in der Schreibkompetenz zurückzuführen sind oder auf andere Varianzquellen. Diese Prüfung ist auch deswegen von besonderer Relevanz, weil für die Elizitierung der Schüler(innen)texte an den Messzeitpunkten jeweils unterschiedliche Parallelversionen der Schreibaufgabe zum Einsatz kamen. Der Kern der Prüfungen besteht in der sukzessiven Testung konfiguraler, metrischer und skalarer Messinvarianz über die Zeit bei zunehmend restriktiven Modellannahmen und in den darauf bezogenen Vergleichen geschachtelter Modelle.

[1] MEZ wird vom BMBF im Rahmen des Schwerpunktes "Sprachliche Bildung und Mehrsprachigkeit" gefördert (Laufzeit 10/2014 bis 9/2019).

2. Das theoretische Konstrukt der Schreibfähigkeit in MEZ

Eine ausführliche Beschreibung des Konstrukts der Schreibfähigkeit findet sich im MEZ-Arbeitspapier Nr. 5 (Usanova und Klinger 2020). Das für die MEZ-Studie maßgebliche Konstrukt der Schreibfähigkeit wurde in seinen Grundlagen im Modellprogramm FörMig (Reich et al. 2009; Gantefort und Roth 2010; Gogolin et al. 2011) entwickelt und richtet sich auf die produktiven Fähigkeiten im Schreiben bei älteren Schüler(inne)n. In Übereinstimmung mit anderen Forschungsarbeiten (Puranik et al. 2008; Wagner et al. 2011) wird davon ausgegangen, dass geschriebene Texte ihrer Natur nach durch mehrere Fähigkeitskomponenten geprägt sind (vgl. Usanova und Klinger 2020). Dementsprechend umfasst auch das in MEZ verwendete, auf Schreibkompetenz bezogene theoretische Modell Teilqualifikationen, die als sprachliche Ressourcen für die schriftliche Produktion von Texten genutzt werden. Nach Ehlich (2005) sind insbesondere diskursive, semantische und morphologisch-syntaktische sowie pragmatische und literale Qualifikationen bedeutsam (vgl. dazu im Einzelnen Usanova und Klinger 2020; vgl. auch Becker-Mrotzek 2014).

Das übergreifende Konstrukt der Schreibfähigkeit der MEZ-Studie setzt sich demnach aus sprachlichen und kognitiven Teilqualifikationen zusammen, die sich auf der Messebene mithilfe von Indikatoren für pragmatische, semantische, morphologisch-syntaktische und literale Qualifikationen konkretisieren lassen (Gantefort und Roth 2010, 582ff.). Bezogen auf die von den Schüler(inne)n produzierten Texte ergibt das Gesamt der betrachteten Indikatoren eine erweiterte Bewertungsgrundlage für die jeweils aktualisierte Textqualität, von der auf die Schreibfähigkeit der Autor(inn)en zurückgeschlossen wird (Klinger et al. 2019; Schnoor 2019).

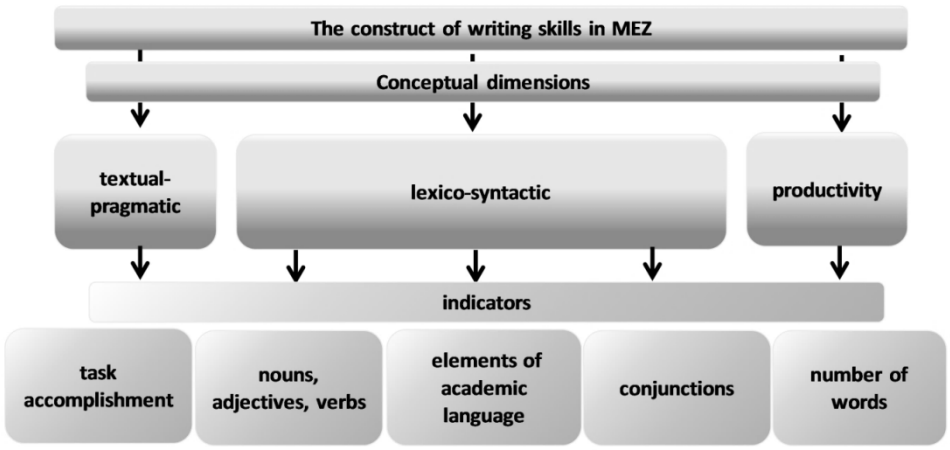


Abbildung 1: Das Indikatorenmodell (aus Usanova & Klinger 2020)

3. Instrumente

Die Messinstrumente für die Erhebung produktiver Schreibfähigkeiten für Deutsch und die Herkunftssprachen sowie für die Schulfremdsprachen wurden auf der Grundlage dieses Modells jeweils in drei Parallelversionen entwickelt und eingesetzt. Jedes Messinstrument besteht aus einer Aufgabenstellung mit dazugehörigen Bildimpulsen sowie aus einem standardisierten Auswertungsbogen.

Die Fähigkeiten der Schüler(innen) im Schreiben wurden mit Parallelformen von Schreibaufgaben getestet, die auf Basis des bestehenden Instruments „Bumerang“ (Dirim und Döll 2009; Reich et al. 2009; Döll 2012; Klinger et al. 2019) entwickelt wurden. Für MEZ wurden auf dieser Basis zwei Parallelversionen („Lebkuchenhaus“ und „Lichterkette“) für Deutsch und die Herkunftssprachen Russisch und Türkisch entwickelt (Usanova und Klinger 2020). Die deutschspra-

chige Version ist Gegenstand des vorliegenden Arbeitspapiers. Für die Schulfremdsprachen Englisch, Französisch und Russisch wurden strukturell vergleichbare, aber thematisch den Curricula für die Fremdsprachen entsprechende Versionen entwickelt.

In den ersten drei Erhebungswellen im MEZ-Projekt wurde pro Messzeitpunkt jeweils eine unterschiedliche Parallelversion verwendet, in der vierten Welle wurden mittels Rotationsverfahren alle drei Aufgaben parallel eingesetzt (IEA Hamburg 2017a, 2017b, 2018a, 2018b). Die als Bildimpulse dienenden Fotografien sind in allen Parallelformen so gestaltet, dass die jeweils dargestellten Motive und Tätigkeiten in allen Versionen vergleichbar komplex sind (Klinger et al. 2019).

Alle drei Versionen des deutsch- und herkunftssprachlichen Instruments enthalten die Aufforderung, einen Zeitschriftenartikel als Arbeitsprobe für eine fiktive Praktikumsbewerbung bei einem Jugendmagazin zu schreiben. Der Artikel soll die Leser(innen) darüber instruieren, wie ein bestimmter Gegenstand hergestellt wird. Dazu sollen die Inhalte von neun Fotografien, die jeweils Arbeitsschritte der Produktion des Gegenstandes darstellen, schriftlich so wiedergegeben werden, dass sie auch ohne Bilder verständlich sind.

Die Aufgabenstellungen waren im Rahmen der MEZ-Erhebungen in den Testheften abgedruckt und wurden zusätzlich von den Testleiter(inne)n vorgelesen bzw. in den Herkunftssprachen als Audiodatei von CD abgespielt. Für die Bearbeitung der Schreibaufgabe wurde jeweils eine einheitliche Testzeit von 30 Minuten vorgegeben. Deutsch-russischsprachigen Proband(inn)en wurde für ihre russischsprachigen Texte ausdrücklich auch die Verwendung lateinischer Schrift gestattet.

Die Analyse der Schülertexte erfolgte für alle Parallelformen in identischer Weise durch geschulte Kodierer(innen) anhand standardisierter Auswertungsbögen, in denen die Indikatoren für Semantik bzw. Textpragmatik, Bildungssprache, Lexik und Syntax nach dem von Reich et al. (2009) vorgelegten Vorschlag (vgl. oben) in Form von sieben konkreten Auswertungskategorien operationalisiert wurden. Hierbei handelt es um folgende Kategorien:

Tabelle 1: Auswertungskategorien der MEZ-Schreibaufgabe für Jugendliche

Aufgabenbewältigung (Punkte):	Der Indikator für die pragmatisch-/semantische Performanz der Texte schätzt die Vollständigkeit bzw. Ausführlichkeit der in der Bildfolge dargestellten Gegenstände und Aktivitäten anhand von neun Inhaltselementen jeweils auf einer vierstufigen Ratingskala ein.
Anzahl verschiedener Nomen (Types):	Der untersuchte Wortschatz umfasst die Anzahl der Types der im Text verwendeten Verben, Nomen und Adjektive.
Anzahl verschiedener Adjektive (Types):	
Anzahl verschiedener Verben (Types):	
Anzahl verschiedener Satzverbindungen (Types):	Als syntaktische Indikatoren werden die für die Strukturierung des Textes und die Herstellung von Textkohärenz wichtigen Satz- bzw. Aussagenverbindungen gezählt.
Anzahl bildungssprachlicher Elemente (Tokens):	Die Verwendung von Bildungssprache in den deutschsprachigen Texten wird mithilfe der Anzahl von Elementen, die für formell-schriftsprachliche Texte charakteristisch sind, bestimmt (Tokens der im Text vorkommenden Nominalisierungen, Komposita, Attributkonstruktionen, Partizipien, Passivkonstruktionen sowie der unpersönlichen Ausdrücke).
Textlänge (Anzahl der Wörter):	Die Textlänge als Maß der *fluency* wird durch die Gesamtzahl der verwendeten Wörter bestimmt, von den Schüler(inne)n durchgestrichene Passagen werden dabei nicht berücksichtigt.

Der Mittelwert (arithmetisches Mittel) der sieben Auswertungskategorien entspricht dem Gesamtscore für die Schreibaufgabe. Da die Werte der einzelnen Auswertungskategorien nicht skaleninvariant sind, wurden die Rohwerte vor der Verrechnung zu einem Gesamtscore zunächst standardisiert. In den MEZ-Datensätzen befinden sich neben den Variablen mit den Rohwerten der einzelnen Auswertungskategorien auch jeweils zwei weitere standardisierte Versionen auf Grundlage der Rohwerte.

z-Werte: Hierfür wurden die Rohwerte der Auswertungskategorien jeweils an der Gesamtstichprobe von MZP1 z-standardisiert [z-Wert = (Rohwert - Mittelwert)/Standardabweichung]. Die resultierenden z-Werte geben für jede der sieben Auswertungskategorien die Leistung einer Person relativ zum mittleren Leistungsniveau der Gesamtstichprobe an. Somit sind die Testwerte zwischen den Auswertungskategorien vergleichbar. Der Mittelwert über die sieben z-standardisierten Variablen ergibt den z-standardisierten Gesamtscore. Dieser hat einen Mittelwert von 0 und eine Standardabweichung von 1, wobei Werte über 0 überdurchschnittliche Leistungen und Werte unter 0 unterdurchschnittliche Leistungen (jeweils in Standardabweichungen) anzeigen.

POMP-Werte (percentage of maximum possible): Hierfür wurden die Rohwerte als Prozentanteil am maximal erreichbaren Wert standardisiert [(Rohwert/maximal erreichbarer Wert) * 100]. Da nur die Aufgabenbewältigung über ein natürliches Maximum verfügt (27 Punkte), wurde für die übrigen Auswertungskategorien der höchste empirische Wert über alle Messzeitpunkte verwendet

4. Datengrundlage

Die hier präsentierten Analysen beruhen auf Daten aller vier Messzeitpunkte der MEZ-Studie. Die Erhebungen für Messzeitpunkt 1 fanden von Januar bis März 2016 (IEA Hamburg 2017a), für Messzeitpunkt 2 von Oktober bis Dezember 2016 (IEA Hamburg 2017b), für Messzeitpunkt 3 von Mai bis Juli 2017 (IEA Hamburg 2018a) und für Messzeitpunkt 4 von Mai bis Juli 2018 (IEA Hamburg 2018b) statt. Beteiligt waren Schüler(innen) an insgesamt 78 Schulen in acht Bundesländern. Berücksichtigt werden im Folgenden diejenigen Schüler(innen), von denen zumindest an einem Messzeitpunkt ein auswertbarer Text vorliegt. In die Messinvarianzprüfungen gehen somit deutschsprachige Texte von insgesamt 2075 Schüler(inne)n ein. Diese verteilen sich auf die Sprachgruppen (Brandt et al. 2019) wie in Tabelle 1 dargestellt.

Tabelle 2: Verteilung der Stichprobe auf die Sprachgruppen

	monolingual deutschsprachig	deutsch-russischsprachig	deutsch-türkischsprachig	anders lebensweltlich mehrsprachig	nicht zuzuordnen*	insgesamt
n	990	364	601	116	4	**2075**
%	47,7	17,5	29,0	5,6	0,2	**100**

45,5 Prozent der Schüler(innen) besuchten einen gymnasialen Bildungsgang, 58,5 Prozent waren weiblich. 50,5 Prozent der Proband(inn)en besuchten zum ersten Messzeitpunkt die Klassenstufe 7 (Kohorte 1), 49,5 Prozent die Klassenstufe 9 (Kohorte 2).

Tabelle 3 gibt einen Überblick über die Mittelwerte und Standardabweichungen der Auswertungskategorien der MEZ-Schreibaufgabe. Die Skalierung der Variablen entspricht den jeweils erreichten individuellen Punktwerten in Relation zur maximal möglichen Punktzahl in Prozent (siehe Abschnitt 3). Die Spannweite der Mittelwerte über die vier Erhebungswellen liegt für die Aufgabenbewältigung zwischen 65,4 Prozent (Welle 1) und 73,8 Prozent (Welle 4), Nomen zwischen 36,1 Prozent (Welle 2) und 52,4 Prozent (Welle 1), für Adjektive zwischen 22,6 Prozent

(Welle 3) und 29,0 Prozent (Welle 2), für Verben zwischen 39,0 Prozent (Welle 3) und 47,1 Prozent (Welle 1), für Satzverbindungen zwischen 30,2 Prozent (Welle 3) und 39,4 Prozent (Welle 1), für Bildungssprachliche Elemente zwischen 24,2 Prozent (Welle 3) und 33,5 Prozent (Welle 1) sowie für die Textlänge zwischen 42,5 Prozent (Welle 1) und 32,2 Prozent (Welle 3).

Die vergleichsweise hohen Mittelwerte der Aufgabenbewältigung gegenüber den anderen Auswertungskategorien resultiert aus ihrem natürlichen Maximum (27 Punkte), das das Auftreten extremer Ausreißer nach oben verhindert. Dieser Mechanismus greift bei den anderen Auswertungskategorien nicht, bei diesen Zählvariablen wurde der höchste gemessene Wert als Maximum gesetzt. Die Aufgabenbewältigung ist auch die einzige Auswertungskategorie, bei der die Mittelwerte von Welle 1 zu Welle 4 kontinuierlich ansteigen. Bei den anderen Auswertungskategorien (bis auf Adjektive) zeigen sich eher ein Trend fallender Mittelwerte von Welle 1 zu Welle 3 und ein Anstieg zu Welle 4. Auf dieses Muster der longitudinalen Mittelwertstruktur der Auswertungskategorien wird später noch genauer eingegangen. Die Reliabilität (Cronbachs α) der Auswertungskategorien beträgt 0,88 (Welle 1), 0,86 (Welle 2), 0,88 (Welle 3), 0,86 (Welle 4).

Tabelle 3: Mittelwerte und Standardabweichungen der Auswertungskategorien (Percentage of maximum possible)

Gesamt n = 2075[2]	**Welle 1 (n=1751)**	**Welle 2 (n=1716)**	**Welle 3 (n=1591)**	**Welle 4**[3] **(n=1085)**
Aufgabenbewältigung (Punkte)	65,4 (14,0)	67.7 (10,6)	70.8 (11,5)	73.8 (11,0)
Nomen (Types)	52,4 (12,9)	36,1 (9,3)	41,8 (10,9)	46,5 (15,1)
Adjektive (Types)	26,7 (15,8)	29,0 (15,5)	22,6 (14,6)	25,8 (16,1)
Verben (Types)	47.1 (13,6)	44.6 (11,8)	39.0 (12,0)	46.2 (15,2)
Satzverbindungen (Types)	39.4 (15,8)	37.8 (15,4)	30.2 (17,0)	32.7 (15,5)
Bildungssprachliche Elemente (Tokens)	33,5 (15,0)	27,0 (12,8)	24,2 (11,6)	27,2 (12,1)
Textlänge (Wörter)	42.5 (12,4)	36.1 (9,2)	32.2 (10,3)	35.2 (12,6)

5. Messinvarianz der Schreibfähigkeit über die Zeit

Der Vergleich identischer Konstrukte über die Zeit setzt faktorielle Invarianz der Messungen voraus. Dies bedeutet, dass Veränderungen in der Varianz-/Kovarianz- und Mittelwertstruktur über die Zeit verlässlich als Veränderungen in den *theoretischen Konstrukten* interpretiert werden können und nicht nur veränderten *Messeigenschaften des Instruments* geschuldet sind. Die longitudinale Messinvarianz gibt somit die Vergleichbarkeit von wiederholten Messungen desselben Konstrukts über die Zeit wieder. In welchem Ausmaß dies möglich ist, wird durch unterschiedliche Grade an Messinvarianz angegeben.

[2] Die hier angegebene Gesamtzahl beinhaltet alle Fälle, für die Informationen vorliegen. Das Programm MPlus, mit dem die unten berichteten Modelle berechnet wurden, erlaubt bei der Schätzung der Modelle die Einbeziehung aller vorliegenden Informationen nach dem FIML-Ansatz (*full information maximum-likelihood*). Dieses Verfahren berechnet eine fallbezogene *likelihood*-Funktion auf der Grundlage der für diesen Fall beobachteten Variablen. Zwar findet keine Imputation der fehlenden Daten statt, gleichwohl gehen die vorhandenen Informationen der unvollständigen Fälle in die Parameterschätzungen ein.

[3] Der erhöhte Ausfall von Teilnehmenden in der 4. Welle geht vor allem auf Schüler(innen) der älteren Kohorte zurück, die aus dem allgemeinbildenden Schulsystem ausschieden oder die Schule wechselten, und ist somit stichprobenbedingt (Heimler 2019, 22).

Für die Messung der Schreibfähigkeit in MEZ bedeutet dies: Will man valide Aussagen über die Entwicklung der Schreibfähigkeit von Schüler(innen) im Zeitverlauf machen, dann muss zunächst sichergestellt werden, dass die Messungen des Konstrukts mittels der MEZ-Schreibaufgabe über die Zeit vergleichbar sind.

Die Analysestrategie für die Überprüfung der longitudinalen Messinvarianz der MEZ Schreibaufgabe im Deutschen besteht aus zwei Schritten: Zunächst wird ein latentes Messmodell spezifiziert, das eine Überprüfung der longitudinalen Messeigenschaften der MEZ-Schreibaufgabe im Deutschen erlaubt (Abschnitt 5.1.). Anschließend wird durch eine sukzessive Auferlegung von Restriktionen der Grad an Invarianz des Messmodells über die Zeit bestimmt (Abschnitt 5.2).

5.1. Spezifikation des Messmodells

Die Konstruktion des Messmodells für die Schreibfähigkeit im Deutschen erfolgt mittels longitudinaler konfirmatorischer Faktorenanalyse (Little 2013). Hierzu werden die oben beschriebenen Auswertungskategorien der MEZ-Schreibaufgabe zu den vier Messzeitpunkten als Indikatoren für die Schätzung eines wiederholt gemessenen latenten Konstrukts der Schreibfähigkeit im Deutschen verwendet (Abbildung 1). Aufbauend auf dem *common factor model* (Thurstone 1947) wird bei konfirmatorischer Faktorenanalyse (Jöreskog 1967, 1969) von der Grundannahme ausgegangen, dass die Messwerte von Personen für direkt gemessene Merkmale (manifeste Indikatoren) kausal auf das Wirken hypothetischer Einflussgrößen auf höheren Abstraktionsebenen (latente Faktoren) zurückführbar sind. Die Beziehung zwischen Indikatoren und Faktoren, d.h. die faktorielle Struktur des Messmodells, wird dabei durch eine Reihe von Modellparametern ausgedrückt, die Auskunft über die Messeigenschaften des Modells geben (zur Einführung in die konfirmatorische Faktorenanalyse siehe Little 2013; Brown 2015; Kline 2016).

Abbildung 2 zeigt eine grafische Darstellung der Spezifikation des longitudinalen Messmodells der Schreibfähigkeit im Deutschen. Das Messmodel besteht zu jedem Messzeitpunkt (t) aus einem latenten Faktor (η_1, η_2, η_3, η_4), der durch vier manifeste Indikatoren gemessen wird: Aufgabenbewältigung (Y1, Y5, Y9, Y13), Verben (Y2, Y6, Y10, Y14), Satzverbindungen (Y3, Y7, Y11, Y15) und Textlänge (Y4, Y8, Y12, Y16). Neben diesen beiden Typen von Variablen, besteht das Modell noch aus einer Reihe von Modellparametern, die sich in zwei Kategorien unterteilen lassen: Modellparameter der Kovarianzstruktur und der Mittelwertstruktur. Zu den Modellparametern der Kovarianzstruktur zählen die die Faktorladungen $\lambda_{i,t}$ als Anteil geteilter Varianz von Indikator und Faktor, die Residuen $\varepsilon_{i,t}$ als indikatorspezifischer Varianzanteil und $\psi_{t,t}$ als Kovarianz zwischen den Faktoren. Zu den Modellparametern der Mittelwertstruktur zählen die latenten Intercepts $\tau_{i,t}$ als modellimplizierte Mittelwerte der Indikatorvariablen und die latenten Mittelwerte der Faktoren αt. Longitudinale Messmodelle haben zudem die Eigenheit, dass dieselben Indikatoren mehrfach im Modell vorkommen (nämlich im Zeitverlauf). Deshalb werden standardmäßig Autokorrelationen der Residualvarianzen der Indikatoren spezifiziert (graue Doppelpfeile). Diese Spezifikationen verbleiben im Modell unabhängig von der jeweiligen Stärke und Signifikanz der Effekte.

In matrizenalgebraischer Schreibweise lässt sich das Messmodell aus Abbildung 2 folgendermaßen darstellen (vgl. Little 2013, 139f.):

$$y_t = T_t + \Lambda_{t\eta t} + \Theta_t \qquad [1]$$

t = bezieht sich auf die jeweilige Erhebungswelle,
y = Messwerte auf den Indikatorvariablen,
T = Spaltenvektor der Mittelwerte der Indikatorvariablen,
Λ = Matrix der Faktorladungen,
η = Messwerte auf den latenten Konstrukten,
Θ = Matrix der indikatorspezifischen Residualvarianzen.

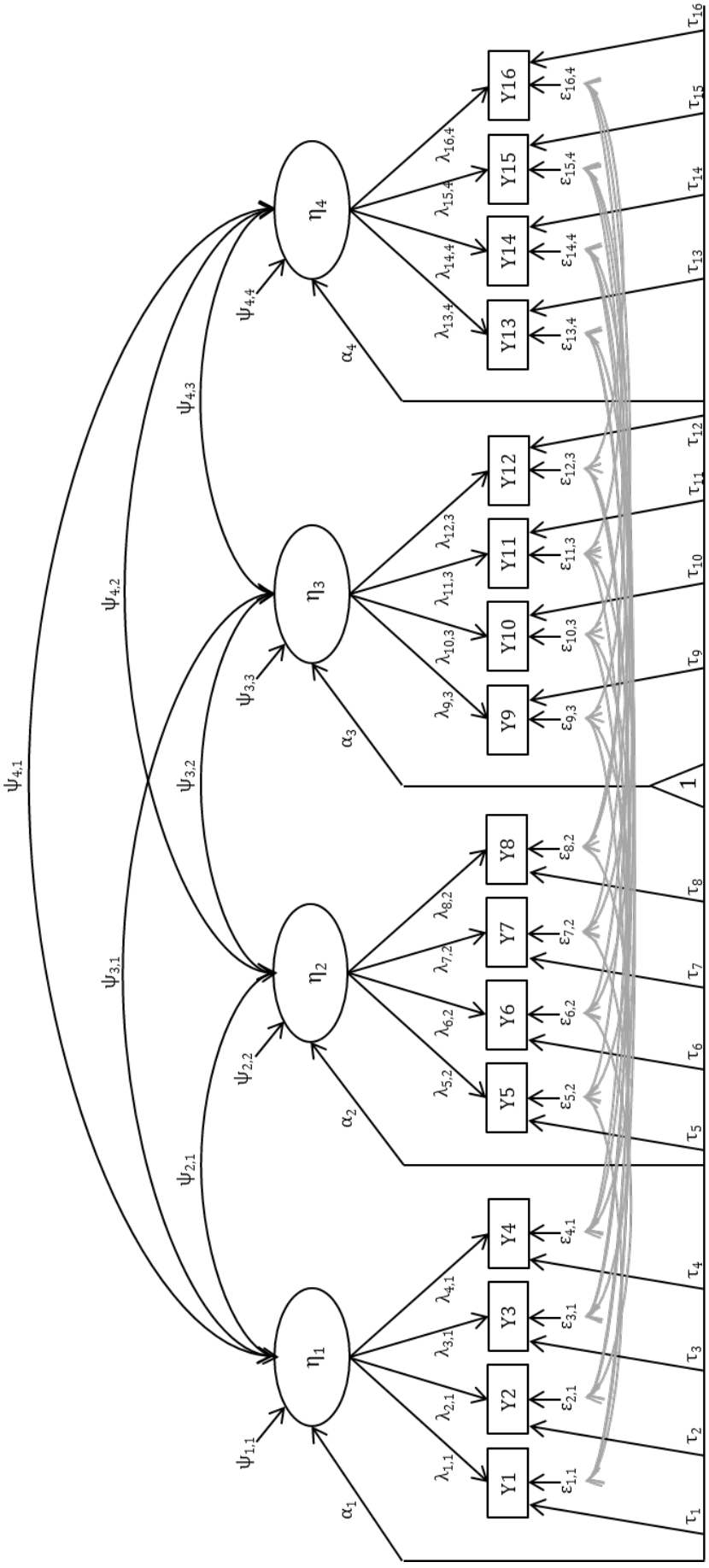


Abbildung 2: Messmodell der Schreibfähigkeiten im Deutschen über vier Erhebungswellen

In der Logik reflektiver Messmodelle sind die Messwerte der manifesten Indikatoren konzipiert als lineare Regression von latentem Konstrukt (η) und Modellparametern (T, Λ, Θ). Die Matrix T_t enthält die (modellimplizierten) Mittelwerte der manifesten Indikatoren zu den vier Erhebungswellen (t). Die Matrix Λ_t enthält die Faktorladungen zwischen manifesten Indikatoren pro Erhebungswelle. Die Matrix Θ_t enthält die indikatorspezifischen Residualvarianzen pro Erhebungswelle.

Die longitudinale Invarianz des Messmodells ist nun abhängig von der Invarianz der Modellparameter über die Erhebungswellen (t). In welchem Ausmaß dies der Fall ist, lässt sich über die Testung von Invarianzhypothesen durch Modellrestriktionen überprüfen.

5.2. Invarianzprüfung des Messmodells

Auf der Ebene von Messmodellen werden vier Grade an Messinvarianz[4] unterschieden (für einen Überblick vgl. Cheung und Rensvold 2002; Wang und Wang 2012; Little 2013; Brown 2015; Kline 2016; siehe Tabelle 4).

1. *Konfigurale Invarianz* (z. B. Horn und McArdle 1992; Meredith 1993) liegt vor, wenn zu allen Messzeitpunkten strukturell gleiche Modelle auf die Daten passen, also dieselbe Anzahl latenter Faktoren mit denselben manifesten Indikatoren assoziiert ist. Ist dies erfüllt, kann angenommen werden, dass das theoretische Konstrukt in den Erhebungswellen in konzeptionell-strukturell vergleichbarer Weise gemessen wurde.
2. *Metrische Invarianz (auch: schwache faktorielle Invarianz*, z. B. Horn und McArdle 1992; Little 1997; Steenkamp und Baumgartner 1998) liegt vor, wenn zusätzlich auch die Faktorladungen der manifesten Indikatoren auf dem latenten Faktor zu den Messzeitpunkten gleich sind (Matrix Λ ist zeitinvariant). Wenn dies erfüllt ist, dann kann angenommen werden, dass das theoretische Konstrukt in allen Erhebungswellen auf derselben Skala gemessen wurde. Dieser Grad an Messinvarianz muss mindestens vorliegen, um die longitudinale Kovarianzstruktur verlässlich analysieren zu können.
3. *Skalare Invarianz (auch: starke faktorielle Invarianz*, z. B. Meredith 1993; Steenkamp und Baumgartner 1998; Little et al. 2007) liegt vor, wenn zusätzlich auch die latenten Intercepts der manifesten Indikatoren zu den Messzeitpunkten gleich sind (Matrix T ist zeitinvariant). Dies bedeutet, dass sich die Werte gleichen, die die manifesten Indikatoren annehmen, wenn der latente Faktor die Ausprägung 0 hat. Wenn dies erfüllt ist, dann haben die Skalen der manifesten Indikatoren in den Erhebungswellen denselben Nullpunkt und dasselbe Intervall. Dieser Grad an Messinvarianz gewährleistet dieselbe operationale Definition der Messskalen und ist notwendig, wenn Analysen der longitudinalen Mittelwertstruktur angestrebt werden.
4. *Residuale Invarianz (auch: strikte faktorielle Invarianz*, z. B. Meredith 1993; Steenkamp und Baumgartner 1998; Little et al. 2007) liegt vor, wenn zusätzlich auch noch die indikatorspezifischen Residualvarianzen zwischen den Messzeitpunkten gleich sind (Matrix Θ ist zeitinvariant). Dies ist ein Hinweis darauf, dass die Reliabilität des Messmodells in allen Teilgruppen gleich ist. Dieser Grad an Messinvarianz ist von nur geringer forschungspraktischer Bedeutung und in den meisten Fällen nicht prüfenswert (Cheung und Rensvold 2002).

Grade an Messinvarianz werden immer für das gesamte Modell geprüft. Wenn ein bestimmter Grad an Messinvarianz für ein Modell nicht bestätigt werden kann, bedeutet dies nicht zwangs-

[4] Die folgenden Ausführungen zu den Graden an Messinvarianz orientieren sich an Schnoor (2019, 98ff.).

läufig, dass hiervon alle Modellteile betroffen sind. Meist handelt es sich um lokale Invarianzprobleme. So wäre zum Beispiel schwache faktorielle Invarianz für ein Modell schon abzulehnen, wenn nur eine einzige Faktorladung nicht invariant ist. Gleiches gilt bei starker faktorieller Invarianz, wenn nur ein latenter Intercept nicht invariant ist. Aus forschungspraktischen Gründen wird in solchen Fällen empfohlen, die Invarianzannahmen zu lockern und die entsprechenden Modellparameter frei zu schätzen. Dies führt dazu, dass der getestete Grad an Messinvarianz zumindest partiell angenommen werden kann. Wenn lediglich ein kleiner Modellteil von dieser Ausnahme betroffen ist und keine gravierenden Verzerrungen zu erwarten sind, dürfen im Fall von partieller schwacher bzw. starker faktorieller Invarianz trotzdem Gruppenvergleiche anhand von Kovarianzen bzw. Mittelwerten vorgenommen werden (Bollen 1989; Byrne et al. 1989; Cheung und Rensvold 2002; Wang und Wang 2012).

Tabelle 4: Grade an Messinvarianz und Prüfbedingungen (Little 2013, eigene Darstellung)

Grad	Bezeichnung	Definition des Testmodells
0	Konfigurale Invarianz	$y_t = T_t + \Lambda_t\eta_t + \Theta_t$
1	Metrische Invarianz (oder schwache faktorielle Invarianz)	$y_t = T_t + \Lambda\eta_t + \Theta_t$
2	Skalare Invarianz (oder starke faktorielle Invarianz)	$y_t = T + \Lambda\eta_t + \Theta_t$
3	Fehlervarianz Invarianz (oder strikte faktorielle Invarianz)	$y_t = T + \Lambda\eta_t + \Theta$

Notiz. Der Buchstabe t in den Modellgleichungen bezieht sich auch den Messzeitpunkt. Mit t gekennzeichnete Matrizen werden für jeden Messzeitpunkt frei geschätzt. Matrizen ohne Kennzeichnung werden mit Gleichheitsrestriktionen über die Zeit belegt.

Das analytische Vorgehen bei der Prüfung longitudinaler Messinvarianz entspricht einer hierarchischen Folge von Hypothesentestungen, in der das Messmodell aus Abbildung 2 (das dem Grad konfiguraler Invarianz entspricht) schrittweise um weiterreichende Invarianzannahmen erweitert wird (Tabelle 2). Dies geschieht mittels Einführung von Restriktionen bei der Schätzung der Modellparameter. Mit jeder weiteren Restriktion wird die Anpassung des Modells an die Daten erschwert, da immer restriktivere theoretische Annahmen mit den Daten in Einklang gebracht werden müssen. Die Grundstruktur des Modells bleibt bei diesen Modifikationen unverändert, sodass eine Abfolge geschachtelter Modelle (*nested models*) entsteht, die jeweils einen bestimmten Grad an Messinvarianz repräsentieren. Welcher Grad an Messinvarianz letztlich angenommen werden kann, wird über Modellvergleiche ermittelt, bei denen das restriktivste Modell angenommen wird, das keine Verschlechterung der Datenanpassung zum vorangegangenen, nächstweniger restriktiven Modell aufweist (Bollen und Long 1993; Cheung und Rensvold 2002; Reinecke 2005; Wang und Wang 2012).

Für die Evaluation der Modellgüte wird als globales Maß der *RMSEA* (*root mean square error of approximation*, Steiger und Lind 1980; Steiger 1998) verwendet. Er testet die approximative Anpassung des Modells an die Daten und ist damit für die Beurteilung komplexer Modelle bei großen Stichproben besser geeignet als der strenge und anfällige χ^2*-Test* (vgl. hierzu Little 1997, 2013). Bei perfekter Modellanpassung gilt *RMSEA* = 0. Werte nahe 0 weisen auf eine annähernde Modellpassung hin. Meist werden für den *RMSEA* folgende Grenzwerte gesetzt: 0 = *perfect fit*; kleiner als 0,05 = *close fit*; 0,05 bis 0,08 = *fair fit*; 0,08 bis 0,10 = *mediocre fit*; größer als 0,10 = *poor fit*. Hu und Bentler (1999) setzen einen *RMSEA* von 0,06 als Grenzwert für einen *good model fit*. Die Ergebnisse des χ^2*-Tests* werden zwar berichtet, aber nicht als Kriterium bei der Beurteilung der Modellgüte verwendet. Als deskriptives Maß der Modellgüte wird der CFI (*comparative fit index,* Bentler 1990) herangezogen, der das spezifizierte Modell mit dessen Nullmodell vergleicht. Der *CFI* kann Werte zwischen 0 und 1 annehmen, wobei Werte nahe 1 eine große Diskrepanz zum Nullmodell anzeigen und auf eine gute Modellpassung hindeuten. Traditionell liegt der Grenzwert für den *CFI* bei 0,90 (Wang und Wang 2012, S. 18). Hu und Bentler (1998, 1999)

empfehlen jedoch, diesen auf 0,95 hochzusetzen. Schermelleh-Engel et al. (2003) sind noch konservativer und empfehlen, ab 0,95 eine akzeptable und ab 0,97 eine gute Modellanpassung anzunehmen.

Das restriktivste Modell, das keine Verschlechterung der Modellanpassung zum vorangegangenen, nächstweniger restriktiven Modell aufweist, wird über Differenzen im CFI ermittelt. Ab einer Differenz von ΔCFI > 0,01 ist eine bedeutsame Verschlechterung in der Modellanpassung anzunehmen und das weniger restriktive Modell definiert den Grad an Messinvarianz (Cheung und Rensvold 2002; Little 2013).

6. Ergebnisse

In Tabelle 5 sind die Ergebnisse der Überprüfung der longitudinalen Messinvarianz der Schreibfähigkeit im Deutschen in MEZ aufgeführt.

Im *Nullmodell* wird Unabhängigkeit der manifesten Indikatoren angenommen; es fungiert als *baseline model*. Zudem gibt es Auskunft über die Gesamtmenge empirischer Informationen, die durch die Varianzen, Kovarianzen und Mittelwerte der Indikatoren für Modellspezifikationen zur Verfügung stehen (144 Freiheitsgrade).

Tabelle 5: Statistiken der Modellanpassung für die Tests auf Invarianzgrade

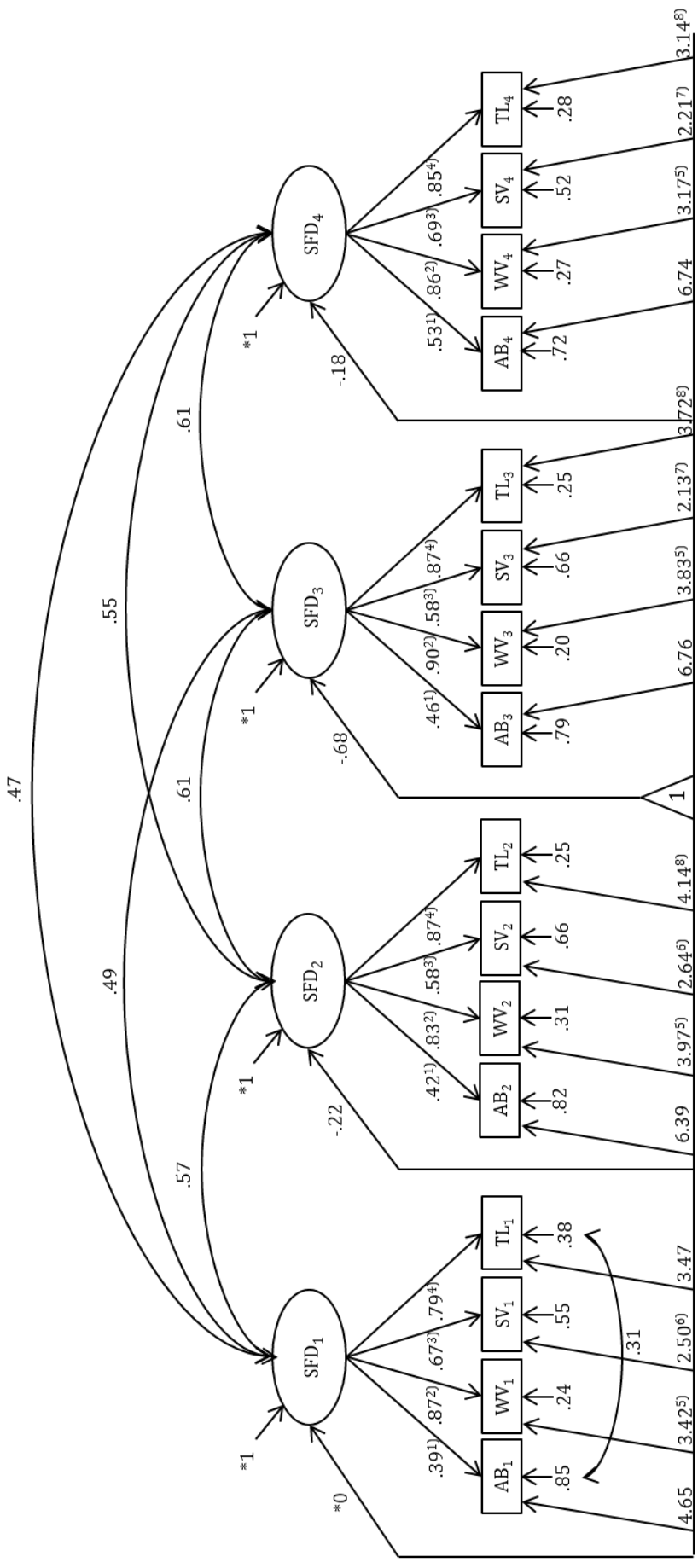
Modell	χ^2	df	p	RMSEA	RMSEA 90% CI	CFI	ΔCFI	Annahme?
Nullmodell	13592.3	144	<.000	-	-	-	-	-
Konfigurale Invarianz	488.7	73	<.000	.052	.048 / .057	.959	-	Ja
Metrische Invarianz	546.9	82	<.000	.052	.048 / .056	.954	.005	Ja
Skalare Invarianz	1837.4	91	<.000	.096	.092 / .100	.828	.126	Nein
Partiell skalare Invarianz (freie Schätzung der Intercepts bei AB & z.T. bei SV)	657.6	86	<.000	.057	.053 / .061	.944	.010	Ja

Das Modell für *konfigurale Invarianz* zeigt mit einer geringfügigen Modifikation der Modellspezifikation eine gute Anpassung an die Daten (χ^2 = 488.7, df = 73, p = <.000; CFI = .959; RMSEA = .052, $CI_{(90\%)}$ = .048/.057).[5] Somit kann angenommen werden, dass die Schreibfähigkeit im Deutschen über die Zeit in konzeptionell-strukturell vergleichbarer Weise gemessen wurde.

Das Modell für *metrische Invarianz* erweist sich ebenso gut vereinbar mit den Daten (χ^2 = 546.9, df = 82, p = <.000; CFI = .954; RMSEA = .052, $CI_{(90\%)}$ = .048/.056). Die Differenz von ΔCFI = 0,005 weist auf keine Verschlechterung der Datenanpassung im Vergleich zum Modell für konfigurale Invarianz hin. Somit kann angenommen werden, dass die Schreibfähigkeit im Deutschen in allen Wellen auf derselben Skala gemessen wurde.

Das Modell für *skalare Invarianz* passt hingegen nicht mehr auf die Daten (χ^2 = 1837.4, df = 91, p = <.000; CFI = .828; RMSEA = .096, $CI_{(90\%)}$ = .092/.100). Dadurch ist nicht gewährleistet, dass die Skalen der latenten Faktoren zu allen Erhebungswellen denselben Nullpunkt haben.

[5] Für die erste Welle wurde eine positive Korrelation der Residualvarianzen bei der Aufgabenbewältigung und der Textlänge zugelassen. Ursächlich für diesen Zusammenhang ist, dass – unabhängig von den Schreibfähigkeiten der Proband(inn)en – die Auswerter(innen) in der ersten Welle längere Texte bei der Aufgabenbewältigung systematisch höher bewertet haben als kürzere Texte.



Modellanpassung:
χ^2 = 657.6, df = 86, p = <.000; CFI = .944; RMSEA = .057, $CI_{(90\%)}$ = .053/.061, Close-Fit p = .003; n = 2075 (full information Likelihood) Schätzer = MLR Koeffizienten = standardized
Legende:
SFD= Schreibfähigkeit Deutsch, AB = Aufgabenbewältigung, WV = Wortschatz: Verben, SV = Satzverbindungen, TL = Textlänge

Abbildung 3: Messmodell der Schreibfähigkeiten im Deutschen mit partieller skalarer Messinvarianz über vier Erhebungswellen (standardisierte Koeffizienten)

Eine Überprüfung der Indikatormittelwerte zeigte, dass diese sowohl bei der Aufgabenbewältigung als auch bei den Satzverbindungen zwischen den Wellen zu stark schwanken. Ein Modell für *partielle skalare Invarianz*, das diesem Umstand Rechnung trägt, zeigt eine zufriedenstellende Anpassung an die Daten (χ^2 = 657.6, df = 86, p = <.000; CFI = .944; RMSEA = .057, $CI_{(90\%)}$ = .053/.061). Hierfür wurden die Intercepts für Aufgabenbewältigung zu allen vier Wellen frei geschätzt sowie die Gleichsetzung der Intercepts der Satzverbindungen nur zwischen Welle 1 und Welle 2 sowie zwischen Welle 3 und Welle 4 vorgenommen.

Abbildung 3 zeigt das Ergebnis der Parameterschätzung für das Modell mit partieller skalarer Messinvarianz in Form von standardisierten Koeffizienten. Da Gleichheitsrestriktionen von Parametern über die Zeit nur auf Ebene der unstandardisierten Koeffizienten sichtbar sind, wurden gleichgesetzte Parameter durch dieselbe hochgestellte Nummerierung gekennzeichnet.

Bezüglich der Stärke der inhaltlichen Assoziation der Faktoren mit den Indikatoren zeigen sich die stärksten Zusammenhänge mit den Verbtypes und der Textlänge, gefolgt von den Satzverbindungen und der Aufgabenbewältigung. Zudem wurde für die erste Welle eine positive Korrelation der indikatorspezifischen Residualvarianzen zwischen Aufgabenbewältigung und Textlänge zugelassen. Das bedeutet, dass diese beiden Auswertungskategorien geteilte Varianz aufweisen, die unabhängig von der Schreibfähigkeit der Jugendlichen ist. Ursächlich hierfür ist eine Überbewertung der Aufgabenbewältigung bei längeren Texten durch die in der ersten Welle eingesetzten Auswerter(innen). Diese Verzerrung wurde durch Nachschulungen in den folgenden Wellen vermieden.

Hinsichtlich der longitudinalen Kovarianzstruktur weisen die Stärken der Zusammenhänge zwischen den Faktoren (.47 bis .61) auf eine relativ hohe Stabilität der Verteilung interindividueller Leistungsunterschiede über die Zeit hin. In Bezug auf die longitudinale Mittelwertstruktur zeigt sich, dass sich der – aufgrund der Mittwertentwicklung der Indikatoren zu erwartende – Abwärtstrend der latenten Mittelwerte der Faktoren zu Welle 2 und Welle 3 bei Welle 4 wieder erholt, aber dennoch unter dem Niveau von Welle 1 verbleibt.

7. Diskussion

Das hier getestete longitudinale Messmodell für Schreibfähigkeiten im Deutschen bezieht sich auf das eingangs vorgestellte theoretische Indikatorenmodell, macht aber nur von vier der sieben Auswertungskategorien Gebrauch. Die Beschränkung von Wortschatzindikatoren auf die Verbtypes vermeidet eine Übergewichtung von Wortschatzphänomenen und ist darüber hinaus messspezifischen Unschärfen zwischen den Wortschatzindikatoren geschuldet. Die Inhaltswortschatztypes der Nomen, Adjektive und Verben korrelieren hoch miteinander, wobei auch gemeinsame Varianzanteile geteilt werden, die nicht auf das hier untersuchte Konstrukt zurückgehen; weiterhin variieren Nomen- und Adjektivtypes deutlicher als die Verbtypes mit der verwendeten Version der Schreibaufgabe (Klinger et al. 2019). Letzteres lässt sich auch für die bildungssprachlichen Elemente annehmen, die sich mit der longitudinalen Messinvarianz nicht vereinbaren lassen, sich gleichwohl in die querschnittlichen Modelle des Gesamtkonstrukts problemlos einfügen.

Die im longitudinalen Messmodell verbliebenen vier manifesten Indikatoren Aufgabenbewältigung, Verbentypes, Satzverbindungen und Textlänge eignen sich für die Schätzung eines wiederholt gemessenen latenten Konstrukts für Schreibfähigkeiten im Deutschen, indem sie das theoretische Konstrukt in den Erhebungswellen in konzeptionell-strukturell vergleichbarer Weise und in allen Erhebungswellen auf derselben Skala messen. Damit ist gewährleistet, dass die longitudinale Kovarianzstruktur verlässlich analysiert werden kann, wie dies z.B. im Rahmen von Panelmodellen geschieht (Selig und Little 2012).

Für Analysen der longitudinalen Mittelwertstruktur ergibt sich allerdings das Problem der Interpretierbarkeit der latenten Mittelwerte über die Zeit. Skalare faktorielle Invarianz kann diesbezüglich für die derzeitige Fassung des Messmodells nicht angenommen werden. Zwar lässt sich ein akzeptables Modell für *partiell*-skalare longitudinale Invarianz schätzen, es zeigen sich jedoch auch dann starke Schwankungen zwischen den latenten Mittelwerten, die von der Interpretation ihrer Entwicklung über die Zeit abraten lassen. Intraindividuelle Veränderungen bzw. deren interindividuelle Differenzen, wie sie z.B. mithilfe latenter Wachstumskurvenmodelle (LGC-Modelle) analysiert werden, sind mit diesen Skalen somit nicht zu erfassen.

Als Begründung für das Verfehlen skalarer Messinvarianz über die Zeit verfolgen wir im Wesentlichen zwei Vermutungen. Zum einen können Unterschiede zwischen den Impulsversionen festgestellt werden, wie sich bei einem ersten Vergleich der am 4. Messzeitpunkt parallel eingesetzten Versionen zeigt. Somit sind die Aufgaben in Bezug auf Schwierigkeit oder Schreibmotivierung offenbar nicht hinreichend ausbalanciert. Zusätzlich lässt sich über die Zeit feststellen, dass die Texte der Schüler(innen) von Welle zu Welle durchschnittlich kürzer werden, was die longitudinale Messung mithilfe von Zählvariablen beeinflusst. Diese Tendenz zeigt sich auch, wenn die Ergebnisse des 4. Messzeitpunktes jeder Aufgabenversion mit denen derselben Version zum jeweils früheren Messzeitpunkt verglichen werden. Hier könnte eine spezifische Form von Paneleffekt eine Rolle spielen, der sich durch die wiederholte Datenerhebung mit strukturell vergleichbaren Schreibaufgaben ergeben haben mag. Möglich wäre z.B. ein Gewöhnungseffekt an den Typ der Aufgabenstellung, der sich auf die Schreibmotivation auswirkt; denkbar ist auch eine zunehmend „ökonomische“ Haltung bei der Aufgabenbearbeitung, die sich in einer sparsameren Verwendung sprachlicher Mittel ausdrückt.

Erfreulich ist, dass sich diese Fluktuation in den longitudinalen Mittelwerten und die damit einhergehende Einschränkung bei der Interpretierbarkeit individueller Entwicklungsverläufe der Schreibfähigkeiten im Deutschen offenbar nicht bedeutsam auf das interindividuelle Leistungsgefüge pro Messzeitpunkt auswirkt. Dies ergibt sich aus der Tatsache, dass die Annahme des Grades metrischer Invarianz aufgrund der hier dargestellten Prüfungen nicht zurückgewiesen werden kann.

Mit den für MEZ entwickelten Aufgaben lassen sich somit im Zeitverlauf Stabilitäten und Veränderungen in individuellen Unterschieden messen und somit Zusammenhangsmuster von Variablen identifizieren. Die genannten Einschränkungen beziehen sich auf die Untersuchung der zeitlichen Dynamik derselben Variable pro Person, d.h. die individuelle Veränderung der betreffenden Variable über die Zeit.

8. Ausblick

Die vorgestellten Analysen geben den gegenwärtigen Stand der Instrumentenentwicklung für die im Projekt MEZ eingesetzten Schreibaufgaben wieder. Wir verfolgen mehrere Strategien, um die festgestellten Probleme im Hinblick auf die Untersuchung longitudinaler Mittelwertstrukturen zu minimieren. Auf der einen Seite können Mittelwertunterschiede zwischen den verschiedenen Versionen der Schreibaufgabe durch geeignete Transformationsprozeduren ausgeglichen werden. Hierbei besteht zum zweiten die zusätzliche Herausforderung, die versionsbezogene Transformation mit einem rechnerischen Ausgleich des vermuteten Paneleffekts zu kombinieren. Dazu prüfen wir z.B. den Einsatz von Gewichtungsverfahren.

Das vorliegende Arbeitspapier beschränkt sich auf die longitudinale Messinvarianzprüfung des Messmodells für Schreibfähigkeiten *im Deutschen*. Eine entsprechende Prüfung der Messmodelle für Schreibfähigkeiten in den anderen Sprachen, die in MEZ untersucht wurden (Herkunftssprachen Russisch und Türkisch sowie in den Schulfremdsprachen Englisch, Französisch und Russisch) wird in einem der folgenden Arbeitspapiere dargestellt.

Literatur

Becker-Mrotzek, M. (2014). Schreibkompetenz. In J. Grabowski (Hrsg.), *Sinn und Unsinn Von Kompetenzen. Fähigkeitskonzepte Im Bereich Von Sprache, Medien und Kultur* (S. 51–71). Leverkusen-Opladen: Barbara Budrich-Esser.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246.

Bollen, K. A. (1989). *Wiley series in probability and mathematical statistics. Applied probability and statistics: Structural equations with latent variables*. New York: Wiley.

Bollen, K. A., & Long, J. S. (1993). *Testing Structural Equation Models*. Newbury Park, CA: SAGE Publications.

Brandt, H., Dünkel, N., & Heimler, J. (2019). Konstruktion und Operationalisierung lebensweltlicher Ein- und Mehrsprachigkeit im Rahmen des Projekts Mehrsprachigkeitsentwicklung im Zeitverlauf (MEZ). *MEZ Arbeitspapiere Nr. 4*. Hamburg. www.mez.uni-hamburg.de.

Brandt, H., Lagemann, M., & Rahbari, S. (2017). Multilingual Development. A Longitudinal Perspective – Mehrsprachigkeitsentwicklung im Zeitverlauf (MEZ). *European Journal of Applied Linguistics, 5*(2).

Brown, T. A. (2015). *Methodology in the social sciences: Confirmatory factor analysis for applied research* (Second edition). New York, London: The Guilford Press.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance. *Psychological Bulletin, 105*(3), 456–466.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233–255.

DESI-Konsortium (Hrsg.) (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie. Ergebnisse der DESI-Studie*. Weinheim [u.a.]: Beltz.

Dirim, I., & Döll, M. (2009). "Bumerang" - Erfassung der Sprachkompetenzen im Übergang von der Schule in den Beruf - vergleichende Beobachtungen zum Türkischen und Deutschen am Beispiel einer Schülerin. In D. Lengyel, H. H. Reich, H.-J. Roth, & M. Döll (Hrsg.), *Von der Sprachdiagnose zur Sprachförderung* (S. 139–146). Münster, New York, NY, München, Berlin: Waxmann.

Döll, M. (2012). *Beobachtung der Aneignung des Deutschen bei mehrsprachigen Kindern und Jugendlichen. Modellierung und empirische Prüfung eines sprachdiagnostischen Beobachtungsverfahrens*. Münster, München [u.a.]: Waxmann.

Ehlich, K. (Hrsg.) (2005). *Bildungsforschung. Bd. 11: Anforderungen an Verfahren der regelmäßigen Sprachstandsfeststellung als Grundlage für die frühe und individuelle Förderung von Kindern mit und ohne Migrationshintegrund*. Bonn, Berlin.

Gantefort, C., & Roth, H.-J. (2010). Sprachdiagnostische Grundlagen für die Förderung bildungssprachlicher Fähigkeiten. *Zeitschrift für Erziehungswissenschaft (ZfE), 13*(4), 573–592.

Gogolin, I., Dirim, I., Klinger, T., Lange, I., Lengyel, D., Michel, U., Neumann, U., Reich, H. H., Roth, H.-J., & Schwippert, K. (2011). *Förderung von Kindern und Jugendlichen mit Migrationshintergrund FörMig. Bilanz und Perspektiven eines Modellprogramms*. Münster, Westf: Waxmann.

Gogolin, I., Klinger, T., Lagemann, M., & Schnoor, B. (2017). Indikation, Konzeption und Untersuchungsdesign des Projekts Mehrsprachigkeitsentwicklung im Zeitverlauf (MEZ). *MEZ Arbeitspapiere Nr. 1*. Hamburg. http://www.pedocs.de/volltexte/2017/14825/pdf/Gogolin_et_al_2017_Indikation_Konzeption_Untersuchungsdesign.pdf.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research, 18*(3-4), 117–144.

Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to under-parameterized model misspecification. *Psychological methods, 3*, 424–453.

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

Hußmann, A., Wendt, H., Bos, W., Bremerich-Vos, A., Kasper, D., Lankes, E.-M., McElvany, N., Stubbe, T. C., & Valtin, R. (Hrsg.) (2017). *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (1. Auflage). Münster: Waxmann.

IEA Hamburg (2017a). *Methodenbericht. MEZ - Mehrsprachigkeitsentwicklung im Zeitverlauf. Erhebung in den Jahrgangsstufen 7 und 9*. 1. Messzeitpunkt - Januar bis März 2016. Hamburg.

IEA Hamburg (2017b). *Methodenbericht. MEZ - Mehrsprachigkeitsentwicklung im Zeitverlauf. Erhebung in den Jahrgangsstufen 8 und 10*. 2. Messzeitpunkt - Oktober bis Dezember 2016. Hamburg.

IEA Hamburg (2018a). *Methodenbericht. MEZ - Mehrsprachigkeitsentwicklung im Zeitverlauf. Erhebung in den Jahrgangsstufen 8 und 10*. 3. Messzeitpunkt - Mai bis Juli 2017. Hamburg.

IEA Hamburg (2018b). *Methodenbericht. MEZ - Mehrsprachigkeitsentwicklung im Zeitverlauf. Erhebung in den Jahrgangsstufen 9 und 11*. 4. Messzeitpunkt - Mai bis Juni 2018. Hamburg.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32*, 443–477.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*, 182–202.

Kline, R. B. (2016). *Methodology in the social sciences: Principles and practice of structural equation modeling* (Fourth edition). New York, London: The Guilford Press.

Klinger, T., Usanova, I., & Gogolin, I. (2019). Entwicklung rezeptiver und produktiver schriftsprachlicher Fähigkeiten im Deutschen. *Zeitschrift für Erziehungswissenschaft (ZfE), 22*(1), 75–103.

Little, T. D. (1997). Mean and Covariance Structures (MACS) Analyses of Cross-Cultural Data. Practical and Theoretical Issues. *Multivariate Behavioral Research, 32*(1), 53–76.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford Press.

Little, T. D., Card, N. A., Slegers, D. W., & Ledford, E. C. (2007). Representing contextual effects in multi-group MACS models. In T. D. Little, J. A. Bovaird, & N. A. Card (Hrsg.), *Modeling contextual effects in longitudinal studies* (S. 121–147). New York: Psychology Press.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543.

Puranik, C. S., Lombardino, L. J., & Altmann, L. J. P. (2008). Assessing the Microstructure of Written Language Using a Retelling Paradigm. *American Journal of Speech-Language Pathology, 17*(2), 107–120.

Reich, H. H., Roth, H.-J., & Döll, M. (2009). Fast Catch Bumerang. Deutsche Sprachversion. Auswertungsbogen und Auswertungshinweise. In D. Lengyel, H. H. Reich, H.-J. Roth, & M. Döll (Hrsg.), *Von der Sprachdiagnose zur Sprachförderung* (S. 209–241). Münster, New York, NY, München, Berlin: Waxmann.

Reinecke, J. (2005). *Strukturgleichungsmodelle in den Sozialwissenschaften*. München: Oldenbourg.

Reiss, K., Weis, M., & Klieme, E. (Hrsg.) (2019). *PISA 2018. Grundbildung im internationalen Vergleich*.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the Fit of Structural Equation Models: Tests of Significance and Deskriptive Goodness-of-Fit Measures. *Methods of Psychological Reasearch Online, 8*(2), 23–74.

Schnoor, B. (2019). *Soziale Herkunft und Bildungssprache. Humankapitalinvestitionen in deutschen, türkischen und vietnamesischen Familien*. Wiesbaden: Springer VS.

Selig, J. P., & Little, T. D. (2012). Autoregressive and cross-lagged panel analysis for longitudinal data. In B. P. Laursen, T. D. Little, & N. A. Card (Hrsg.), *Handbook of developmental research methods* (S. 265–278). New York, NY: Guilford Press.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research, 25*(1), 78–107.

Steiger, J. H. (1998). A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modeling: A Multidisciplinary Journal, 5*(4), 411–419.

Steiger, J. H., & Lind, J. C. (1980). *Statistically based tests for the number of common factors*. unveröffentlichtes Manuskript zu einer Präsentation dem Psychometric Society Annual Meeting in Iowa City, USA.

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago: University of Chicago Press.

Usanova, I., & Klinger, T. (2020). Die Messung mehrsprachiger Schreibfähigkeiten im Projekt Mehrsprachigkeitsentwicklung im Zeitverlauf: Schreibaufgaben, Kodierung und Skalen. *MEZ Arbeitspapiere Nr. 5*. Hamburg. www.mez.uni-hamburg.de.

Wagner, R. K., Puranik, C. S., Foorman, B., Foster, E., Wilson, L. G., Tschinkel, E., & Kantor, P. T. (2011). Modeling the development of written language. *Reading and Writing, 24*(2), 203–220.

Wang, J., & Wang, X. (2012). *Wiley Series in Probability and Statistics: Structural Equation Modeling. Applications Using Mplus*. Sussex, UK: Wiley.

Wendt, H., Bos, W., Selter, C., Köller, O., Schwippert, K., & Kasper, D. (Hrsg.) (2016). *TIMSS 2015. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster, New York: Waxmann.