

Empirische Sonderpädagogik, 2014, Nr. 4, S. 300-312
ISSN 1869-4845 (Print) · ISSN 1869-4934 (Internet)

Testtheoretische Absicherung der ESGRAF 4-9

Christian Rietz & Hans-Joachim Motsch

Universität zu Köln

Zusammenfassung¹

Gängige Sprachentwicklungstests überprüfen den Erwerbsstand grammatischer Regeln bei Kindern weder umfassend noch therapierelevant. Mit dem Testverfahren ESGRAF-R bzw. dessen Weiterentwicklung ESGRAF 4-9 soll diesem Defizit an einschlägigen diagnostischen Verfahren begegnet werden. Im Rahmen der vorliegenden empirischen Studie, die Bestandteil des Forschungsprojektes GED 4-9 (Grammatikerwerb deutschsprachiger Kinder zwischen 4 und 9 Jahren) ist, wurde die ESGRAF 4-9 testtheoretisch unter Berücksichtigung der Aspekte Objektivität, Reliabilität und Validität (Inhaltsvalidität sowie divergente und konvergente Konstruktvalidität) abgesichert. Die guten Befunde werden ergänzt durch Hinweise darauf, dass eine anstehende Normierung der ESGRAF 4-9 in Bezug auf das Alter erfolgen sollte.

Schlüsselwörter: Sprachdiagnostik, grammatische Störungen, psychometrische Testverfahren, Spezifische Sprachentwicklungsstörung

The Psychometric Quality of the ESGRAF 4-9

Abstract

Commonly used assessments of language development usually do not assess children's acquisition of grammatical rules extensively nor do they assess them with regard to therapeutic relevance. ESGRAF-R and its enhanced version ESGRAF 4-9 aim at closing this lack of appropriate assessment tools. In this empirical study, which is part of the research project GED 4-9 (acquisition of grammatical rules by German-speaking children aged 4 to 9), ESGRAF-R is being analyzed with regard to quality standards of test and measurement theory. Therefore, objectivity, reliability and validity (content validity as well as divergent and convergent construct validity) are taken into consideration. The results show good quality standards for ESGRAF-R and suggest that future research should focus on determining age-related standard values.

Key words: linguistic assessment, morphosyntactic deficits, psychometric profiling, specific language impairment

¹ Für die Schulung des Testteams und die logistische Betreuung der Studie danken wir der Projektkoordinatorin des GED 4-9 Projektes, Frau M.A. Lena Becker. Für die Mithilfe bei der Endredaktion des Manuskripts danken wir Frau M.Sc. Simone van Koll.

Der Erwerb grammatischer Strukturen deutschsprachiger Kinder wird entgegen früherer Annahmen von Linguistinnen und Linguisten (Szagun, 2007) erst im Schulalter abgeschlossen. Während die frühen grammatischen Regeln (syntaktische Verbzweitstellungsregel im Hauptsatz, morphologische Subjekt-Verb-Kontroll-Regel) von den meisten Kindern bereits im dritten Lebensjahr korrekt angewendet werden, ist der Erwerb der späten grammatischen Regeln (Kasusreaktion, Nebensatzregel) bei vielen Kindern beim Eintritt in die Schule noch nicht abgeschlossen (Motsch, 2013).

Die in den letzten Jahren publizierten Sprachentwicklungstests „Patholinguistische Diagnostik bei Sprachentwicklungsstörungen“ (PDSS; Kauschke & Siegmüller, 2010), „Potsdam-Illinois Test für Psycholinguistische Fähigkeiten“ (P-ITPA; Esser, Wyszkon, Balaschik & Hänsch, 2010), „Sprachstandserhebungstest für Kinder im Alter zwischen 5 und 10 Jahren“ (SET 5-10; Petermann, 2010) und „Entwicklungstest Sprache für Kinder von vier bis acht Jahren“ (ETS 4-8; Angermaier, 2007) überprüfen den Erwerbsstand grammatischer Regeln des Kindes lediglich stichprobenmäßig anhand ausgewählter „grammatischer Marker“. Beliebte ist die effiziente Überprüfung der Pluralformen des Deutschen. Dies erscheint bereits theoretisch fragwürdig, da die Pluralmorphologie nicht regelgeleitet erworben wird. Empirisch zeigt sich die Fragwürdigkeit der präferierten Pluralitems durch die fehlende interne Konsistenz und fehlende Korrelationen zwischen Pluralsubtests. Einige Testverfahren überprüfen lediglich morphologische, aber nicht syntaktische Fähigkeiten (ETS 4-8) oder aggregieren Aussagen über unterschiedliche grammatische Strukturen in einem Subtest (P-ITPA: Präteritum, Perfekt, Komparationsformen und Kasus) zu einem einzigen Rohpunktwert. Verfahren, die Kasus getrennt überprüfen, verwenden eine zu geringe Itemzahl (PDSS: Akkusativ 2(4), Dativ 5). Somit erhält die Diagnostikerin bzw. der Diagnostiker durch die Ergebnisse grammatischer Subtests nicht die therapieplanungsrelevante Frage beantwortet, auf welchem Er-

werbsstand das getestete Kind bezüglich der einzelnen wichtigen Regeln des Deutschen steht.

Das Testverfahren ESGRAF 4-9

Mit dem erstmals 2009 publizierten und 2013 in überarbeiteter Version erschienenen modularisierten Verfahren „Evozierte Sprachdiagnose grammatischer Fähigkeiten“ (ESGRAF-R; Motsch, 2013) war es möglich, wichtige grammatische Regeln und Phänomene umfassend und hypothesengeleitet zu überprüfen. Kritik erfuhr das teilstandardisierte spieldiagnostische Verfahren aufgrund der fehlenden testtheoretischen Absicherung und Normierung. Die erreichten Ergebnisse waren therapierelevant, entsprachen aber nach strengen psychometrischen Kriterien lediglich der Orientierungsgröße eines elaborierten Screenings. Im Zuge der Weiterentwicklung von der ESGRAF-R zur ESGRAF 4-9 erfolgte nicht nur eine Standardisierung, sondern auch eine Erweiterung. Die Überprüfung elementarer grammatischer Regeln wurde im Hinblick auf den Einsatz im Schulalter ergänzt durch die Überprüfung formal komplexerer und kognitiv anspruchsvoller Strukturen (Passiv- und Konjunktivsätze, Genitiv).

Die vorliegende empirische Voruntersuchung, die das Ziel der testtheoretischen Absicherung der ESGRAF 4-9 hat, entstand im Rahmen des multizentrischen Forschungsprojektes „Grammatikerwerb deutschsprachiger Kinder zwischen 4 und 9 Jahren“ (GED 4-9). Ziel dieses Vorhabens ist es, die grammatischen Fähigkeiten von mehr als 1000 Kindern zu erheben und das Verfahren zu normieren.

Die testtheoretische Absicherung der ESGRAF 4-9

Im Folgenden soll die testtheoretische Absicherung der ESGRAF 4-9 im Rahmen einer empirischen Untersuchung unter Berücksichtigung der Testgütekriterien

– Objektivität,

- Reliabilität und
- Validität

dargestellt werden. Die testtheoretische Absicherung geschieht auf Grundlage der Klassischen Testtheorie (KTT)² (vgl. zusammenfassend zur KTT Moosbrugger, 2012b).

Methode

Untersuchungszeitraum und -ort

Die Untersuchung, die das Ziel der testtheoretischen Absicherung der ESGRAF 4-9 hat und einer größeren Erhebung im Rahmen des Forschungsprojektes GED 4-9 zur Normierung des Verfahrens vorgeschaltet ist, wurde im Zeitraum zwischen dem 16. September 2013 und dem 30. September 2013 an Kindertagesstätten und Schulen in Köln bzw. im Kölner Umland durchgeführt. Insgesamt wurden Daten an $N = 113$ monolingual deutschsprachigen Kindern erhoben (vgl. auch Motsch & Becker, 2014).

Untersuchungsdesign und -variablen

Das Untersuchungsdesign sah eine Quotierung der Stichprobe nach den Merkmalen Alter und Geschlecht vor (siehe Tabelle 2), da aufgrund von theoretischen Vorannahmen zu erwarten war, dass diese Variablen einen Einfluss auf die Testergebnisse haben.

Weiterhin wurde als unabhängige Variable berücksichtigt, ob die Kinder aus dem städtischen oder dem ländlichen (suburbanen) Raum stammen. Es erfolgte die Erhebung zusätzlicher soziodemografischer Variablen, deren Einfluss auf die ESGRAF 4-9 es im Rahmen der Untersuchung zu überprüfen galt: grammatische Kompetenz (erfasst durch Einschätzungen durch die Erzieherinnen und Erzieher bzw. die Lehrkräfte) und etwaige Sprachauffälligkeiten bei den Kindern und Bildungsabschluss der Mutter. Zur Validierung der ESGRAF 4-9 wurden Untertests aus dem Entwicklungstest Sprache 4–8 Jahre (ETS 4-8, Angermaier, 2007), dem Sprachstandserhebungstest für Fünf- bis Zehnjährige (SET 5-10, Petermann, 2010), dem Inventar diagnosti-

Tabelle 1: Testaufbau der aus sechs Untertests mit 212 Items bestehenden ESGRAF 4-9

Testteil	Inhalt	Itemanzahl
Subtest 1a	Verbzweitstellungsregel im Hauptsatz	45 Items
Subtest 1b	Subjekt-Verb-Kontrollregel	42 Items
Subtest 2	Verbendstellungsregel im subordinierten Nebensatz	20 Items
Subtest 3	Genussicherheit	20 Items
Subtest 4a	Akkusativ in Nominal- und Präpositionalphrasen	27 Items
Subtest 4b	Dativ in Nominal- und Präpositionalphrasen	30 Items
Subtest 5	Pluralmarkierung der neun Pluralformen des Deutschen	9 Items
Subtest 6	Späte grammatische Fähigkeiten (Passiv- und Konjunktivkonstruktionen, Genitiv)	19 Items

Anmerkungen. Subtest 3 überprüft die eigentlich lexikalische Fähigkeit der sicheren Genuszuordnung an den Nomen, die im Subtest 4 zur Kasusüberprüfung im Akkusativ und Dativ Verwendung finden. Dies erlaubt eine qualitative Interpretation der Kasusergebnisse im Hinblick auf den Einfluss möglicher Genusunsicherheit.

² Die testtheoretische Absicherung mit probabilistischen Modellen wurde im Rahmen der Testentwicklung diskutiert. Gerade in Bezug auf die Voraussetzung der Itemhomogenität (vgl. Moosbrugger, 2012a, S. 228-232) gehen die Autoren davon aus, dass diese Annahme zum jetzigen Zeitpunkt der Testkonstruktion noch „zu streng“ ist. Eine Erweiterung der testtheoretischen Absicherung bzw. eine Ergänzung um die Überprüfung mit Modellen der Item-Response-Theorie (IRT) wird aber längerfristig nicht ausgeschlossen.

Tabelle 2: Untersuchungsstichprobe (quotiert nach Alter und Geschlecht)

	Alter (Monate)										Gesamt
	48-53	53-59	60-65	66-71	72-77	78-83	84-89	90-95	96-101	102-107	
Jungen	6	4	7	5	5	4	6	7	4	5	53
Mädchen	5	9	5	8	7	4	6	6	5	5	60
Gesamt	11	13	12	13	12	8	12	13	9	10	113

scher Informationen bei Sprachentwicklungsauffälligkeiten (IDIS, Schöler, 1999) sowie der Kaufmann-Assessment Battery for Children (K-ABC, Melcher & Preuß, 2009) erhoben.

Untersuchungsdurchführung

Das Team, das die Studie zur testtheoretischen Absicherung der ESGRAF 4-9 durchführte, bestand aus sieben Diagnostikerinnen und Diagnostikern, die mittels eines zweiteiligen Workshops eine Einweisung in den Aufbau und die Durchführung des Verfahrens bzw. der Validitätstests und die Auswertung der erhobenen Daten erfahren hatten. Um das Verständnis des Kontextes der Studie zu sichern und die Testerinnen und Tester für die Erhebung grammatischer Strukturen zu sensibilisieren, wurde im ersten Schritt theoretisches Wissen zum normgerechten bzw. gestörten Erwerb grammatischer Fähigkeiten und dem Aufbau der eingesetzten Testverfahren vermittelt. Darauf nahmen sie an einer Videoschulung teil, in deren Rahmen sie sich Filmsequenzen von Testsituationen ansahen, um die Durchführung des Verfahrens und die Interaktion zwischen Diagnostikerin bzw. Diagnostiker und Kind während der Testung vorab einmal beobachten zu können. Anschließend übten die Diagnostikerinnen und Diagnostiker das Verfahren per Rollenspiel unter Supervision und führten es eigenständig mit Kindern aus dem privaten Umfeld durch. Weiterhin nahmen die Testleiterinnen und Testleiter an einem Workshop zur Auswertung der erhobenen Daten mit Fokus auf dem Aufbau des ESGRAF-Protokollbogens sowie auf dem Transfer der per Audioaufnahme erhobenen Daten in den Aus-

wertungsbogen teil. Anhand ausgewählter Aufnahmen von Testungen wurde die Auswertung exemplarisch durchgeführt und vertiefend geübt.

Die Erhebung der grammatischen Fähigkeiten erfolgte in Einzelsettings in Räumlichkeiten der jeweiligen Einrichtung. Das Kind wurde in das Rahmenthema „Zirkus“ eingeführt und durch die verschiedenen, spielerisch eingebetteten Subtests geleitet. Die Dauer der Erhebungen betrug ca. 90 Minuten. In den Kindertagesstätten waren für die Durchführung des Verfahrens drei Termine à 30 Minuten anberaumt, in den Schulen zwei Schulstunden. Während der Testungen erfolgte eine Aufzeichnung der evozierten Kindesäußerungen per Audiogerät. Anschließend wurden die Audioaufnahmen abgehört, um die erhobenen Daten in den Protokollbögen zu übertragen.

Weitere Erhebungen

Zusätzlich zur Erhebung der Tests an den $N=113$ Kindern wurden bei $n=20$ Kindern Doppelauswertungen vorgenommen, die gleichmäßig auf die Jahreskohorten der Stichprobe (4;0 – 4;11, 5;0 – 5;11, 6;0 – 6;11, 7;0 – 7;11, 8;0 – 8;11) verteilt waren. Von Seiten der beiden Autoren dieser Studie bemühte man sich um eine Gleichverteilung von Jungen und Mädchen. Die Erst-Auswertung der Daten erfolgte durch drei verschiedene Tester. Eine Diagnostikerin des Teams nahm anschließend die Doppelauswertungen vor. Bei Kindern aus einer weiteren Stichprobe der Größe $n=30$ werden im Herbst 2014 Wiederholungsbefragungen mit der ESGRAF 4-9 durchgeführt.

Beschreibung der Untersuchungstichprobe

Das Durchschnittsalter der im September 2013 untersuchten $N=113$ Kinder zwischen 4 und 9 Jahren betrug 76.58 Monate. Wie Tabelle 2 zu entnehmen ist, wurden $n=53$ Jungen (46.9 %) und $N=60$ Mädchen (53.1 %) untersucht. $N=83$ Kinder (73.5 %) stammten aus einer Kindertagesstätte/Schule im städtischen Raum, $N=30$ Kinder (26.5 %) aus einer Einrichtung im suburbanen/ländlichen Raum. Bezüglich des Bildungsniveaus der Eltern (genauer der Mutter) hatten $N=9$ (8.0 %) Kinder Mütter mit einem Hauptschulabschluss, $N=32$ (28.3 %) Kinder Mütter mit einem Realabschluss und $N=62$ (73.7 %) Kinder Mütter mit einem Fachhochschulabschluss oder Abitur. Bei $N=74$ Kindern (65.5 %) lag keine Sprachauffälligkeit vor, als sprachauffällig wurden $N=39$ (35.4 %) Kinder klassifiziert. Gemäß der Einschätzung der Erzieherinnen und Erzieher bzw. der Lehrkräfte waren die grammatischen Fähigkeiten der Kinder bei $N=33$ (29.2 %) sehr gut, bei $N=57$ (50.4 %) gut, bei $N=21$ (18.6 %) befriedigend und bei $N=2$ (1.8 %) ausreichend.

Ergebnisse

Objektivität

Unter dem Oberbegriff Objektivität werden die Durchführungs-, die Auswertungs- und die Interpretationsobjektivität zusammengefasst (vgl. z. B. Lienert & Raatz, 1998).

Eine hohe Durchführungsobjektivität ist gegeben, wenn der Test möglichst standardisiert ist. Von einer optimalen Standardisierung (vgl. hierzu auch Moosbrugger & Kelaiva, 2012a) kann ausgegangen werden, wenn die Testperson in der Testsituation die einzige Variationsquelle darstellt. Die Standardisierung von Testmaterial und Untersuchungsdurchführung der ESGRAF 4-9 folgt dieser Forderung. Die Variationsquelle „Untersuchungsleiterin bzw. Untersuchungsleiter“, die bei allen Verfahren, die von einer menschlichen Person durchgeführt werden, weitere Varianzanteile generieren kann, wurde im Rahmen der Durchführung der ESGRAF 4-9 durch eine ausführliche Untersucherschulung möglichst reduziert.

Auswertungsobjektivität ist gegeben, wenn das Ergebnis der ESGRAF 4-9 nicht von der Person des Testauswerters abhängt. Die Auswertung der ESGRAF 4-9 erfolgt erst nach Abschluss der Testsituation. Daher steht bei der Überprüfung der Auswertungsobjektivität die Frage im Vordergrund, ob verschiedene Auswerter eines Testprotokolls zu idealerweise identischen Ergebnissen kommen. Nach Moosbrugger und Kelava (2012a) kann die Auswertungsobjektivität über den Grad der Übereinstimmung gemessen werden, in dem zwei oder mehr Testauswerter zu übereinstimmenden Ergebnissen kommen. Als Maße für die Beobachterübereinstimmung lässt sich im vorliegenden Fall bei als intervallskaliert anzunehmenden Summenscores der Intraklassenkorrelationskoeffizient (ICC) heranziehen (zum Intraklassenkorrelationskoeffizien-

Tabelle 3: Intraklassenkorrelationskoeffizienten gemäß dem Modell ICC(2,1) justiert und unjustiert für die Auswertung der ESGRAF 4-9 mit zwei Ratern bei insgesamt $N=20$ Kindern

	Kind 1	Kind 2	Kind 3	Kind 4	Kind 5	Kind 6	Kind 7	Kind 8	Kind 9	Kind 10
ICC justiert	0.938	0.833	0.750	0.723	0.986	0.925	0.901	0.875	0.985	0.914
ICC unjustiert	0.937	0.828	0.751	0.730	0.986	0.927	0.888	0.879	0.984	0.913
	Kind 11	Kind 12	Kind 13	Kind 14	Kind 15	Kind 16	Kind 17	Kind 18	Kind 19	Kind 20
ICC justiert	0.888	0.931	0.981	0.950	0.959	0.934	0.973	1.000	0.938	0.918
ICC unjustiert	0.887	0.927	0.981	0.946	0.953	0.935	0.974	1.000	0.939	0.916

ten vgl. die auch heute noch aktuelle Arbeit von Asendorpf & Wallbott (1979), McGraw & Wong (1996a, 1996b) oder Wirtz & Caspar (2002)). Die Berechnung der Intraklassenkorrelationskoeffizienten erfolgt varianzanalytisch. Das hier gewählte Modell ICC(2.1) wird unter zwei Analysefragestellungen angewendet: Zum einen wird ein justierter Intraklassenkorrelationskoeffizient berechnet, der die relative Ähnlichkeit der Urteile im Sinne eines Korrelationskoeffizienten abbildet. Zum zweiten wird aber auch unter der strengeren Prämisse eines unjustierten Koeffizienten geprüft, ob nicht nur die relativen Positionen, sondern auch die absoluten Positionen der Bewertungen ähnlich sind^{3,4}.

Der über alle Bewertungen berechnete ICC beträgt justiert 0.925 und unjustiert 0.924. Insgesamt kann somit von einer sehr guten Auswertungsobjektivität ausgegangen werden.

Die Interpretationsobjektivität liegt vor, wenn die Anwender eines Verfahrens aufgrund eines Testergebnisses zu derselben Schlussfolgerung kommen. Bereits bei der ESGRAF-R lagen für alle überprüften grammatischen Regeln für die Gruppe der Kinder im Einschulungsalter (5;8-7;4 Jahre) Prozenträge (PR) vor, die erlaubten, Therapieziele zu begründen, sobald ein Kind einen PR unter 15 erreicht hat (age-related-model). Wurde der PR 15 bei mehreren grammatischen Regeln unterschritten, erfolgte die Bestimmung des prioritären Therapieziels entwicklungsorientiert auf dem Hintergrund des Wissens über die normale Erwerbsreihenfolge der grammatischen Strukturen (skill-related-model, Lees & Urwin, 1998). Das Kriterium

der Entwicklungsorientierung ermöglicht, die scheinbare Dichotomie zwischen dem quantitativen Altersspannen-Modell und dem qualitativen Lernschrittmodell zu überwinden (Motsch, 2010).

Da für die ESGRAF 4-9 im Rahmen der laufenden Normierungsstudie Normwerte ermittelt werden können, sind die Testergebnisse unter Rekurs auf diese Normwerte in Zukunft für den gesamten Altersrange zwischen 4;0 und 8;11 Jahren in Halbjahresschritten eindeutig interpretierbar.

Reliabilität

Die Reliabilität untersucht die Messgenauigkeit eines Verfahrens. Bei der testtheoretischen Absicherung der ESGRAF 4-9 wurden als Reliabilitätsmaße die Testhalbierungsmethode (Split-Half-Methode) sowie die Bestimmung der internen Konsistenz verwendet (vgl. hierzu die sehr anschauliche Darstellung in Schermelleh-Engel & Werne, 2012). Bei der Testhalbierungsmethode wird ein Test, beispielsweise der aus 45 Items bestehende ESGRAF 4-9-Subtest 1 Morphologie: Subjekt-Verb-Kongruenz, in zwei möglichst gleich große Hälften geteilt. Anschließend sind die beiden Hälften miteinander zu korrelieren. Die interne Konsistenz, die auf Cronbach (1951) zurückgeht, kann quasi als Generalisierung der Testhalbierungsmethode auf so viele Testteile wie Items verstanden werden. Der aus der Berechnung resultierende Koeffizient Cronbachs α lässt sich varianzanalytisch herleiten und konzeptuell als mittlere Korrelation zwischen allen Items der jeweiligen Skala verstehen⁵.

³ Die Berechnung der Intraklassenkorrelationskoeffizienten mit SPSS erfolgt über die Reliabilitätsanalyse. Im Menü der Reliabilitätsanalyse kann unter „Statistiken“ der Intraklassenkorrelationskoeffizient entsprechend dem jeweiligen Modell justiert bzw. unjustiert angefordert werden.

⁴ Dass die Intraklassenkorrelationskoeffizienten bei der „strengeren“ unjustierten Lösung manchmal etwas besser sind als bei der liberaleren justierten Lösung, ist darauf zurückzuführen, dass sich durch die Justierung die Quadratsummen und Varianzanteile etwas ändern.

⁵ Im Fall von dichotomen Items (wie bei den meisten Skalen der ESGRAF 4-9) wird α aber nicht nach der Formel von Cronbach berechnet, sondern geht auf eine etwas ältere Kuder-Richardson Formel zurück, die sich allerdings von α nur geringfügig unterscheidet (vgl. Amelang & Schmidt-Atzert, 2006, S. 146). SPSS berechnet α im Fall von dichotomen Items nach der Formel von Kuder und Richardson und im Fall polytomer Items nach der Formel von Cronbach.

Tabelle 4: Reliabilitäten, Itemschwierigkeiten und Subskalenmittelwerte/-standardabweichungen

Testteil	Items	Cronbachs α	Split-Half- Koeffizient	Leichtestes Item	Schwerstes Item	Skala Mittel- wert	Skala Stdabw
Frühe Fähigkeiten							
Subtest 1a Morphologie: Subjekt-Verb-Kon- gruenz	45	0.91	0.39	0.99	0.54	0.85	0.16
Subtest 1b Syntax: Verbzweit- stellungsregel	42	0.90	0.45	0.99	0.55	0.87	0.14
Subtest 3 Genus (Lexikon)	20	0.91	0.82	0.91	0.68	0.84	0.21
Späte Fähigkeiten							
Subtest 2 Syntax: Verbend- stellungsregel	20	0.79	0.73	0.97	0.70	0.86	0.14
Subtest 4a Morphologie: Dativ	30	0.93	0.93	0.96	0.14	0.60	0.26
Subtest 4b Morphologie: Akkusativ	27	0.79	0.64	0.97	0.12	0.75	0.15
Subtest 5 Morphologie: Plural	9	0.38	0.47	0.99	0.83	0.94	0.10
Subtest 6 Syntax: Passiv	5	0.64	0.53	1.49 (0.75)	0.49 (0.24)	0.87	0.51
Subtest 6 Morphologie: Kasus: Genitiv	8	0.87	0.79	1.49 (0.75)	0.49 (0.24)	1.02	0.57
Subtest 6 Konjunktiona- sätze	6	0.96	0.94	0.91 (0.45)	0.82 (0.41)	0.87	0.89
Subtest 6 Gesamtscore	19	0.88	0.60	1.49 (0.75)	0.49 (0.24)	0.93	0.43

Anmerkungen. Die Berechnung der Kennziffern für das leichteste und schwerste Item basiert auf der Proportion der Richtiglöser an den Gesamtlösern (dichotome Items bei den Subtests 1 bis 5). Bei Subtest 6 liegen dreifach gestufte Items vor. Hier sind die Kennziffern in Klammern auf den Bereich zwischen 0 und 1 normiert und können so analog zu den Kennziffern der anderen Subskalen interpretiert werden. Die Werte unter „Skala Mittelwert“ stellen den Mittelwert der Items der jeweiligen Skala dar, mit „Skala StdAbw“ ist die Standardabweichung der Skalenwerte bezeichnet.

Bis auf Subtest 5 (Plural) liegen die Schätzungen der internen Konsistenz sowie der Split-Half-Reliabilitäten für die gesamten Subskalen in einem akzeptablen bis sehr guten Bereich (vgl. zu einer Einschätzung der Höhe der Koeffizienten die Diskussion in Schermelleh-Engel & Werner, 2012).

Inhalts- und Konstruktvalidität

Eine hohe Inhaltsvalidität eines Verfahrens liegt vor, wenn die Inhalte eines Tests tatsächlich das interessierende Merkmal erfassen, wobei sich die Inhalte hier vor allem auf das Stimulusmaterial und die Antwortmöglichkeiten des Verfahrens beziehen (vgl. zu den verschiedenen Formen der Validität Hartig, Frey & Jude, 2012). Aufgrund der langjährigen Entwicklung der ESGRAF 4-9, aufgrund der Erfahrungen mit der ESGRAF (Motsch, 2000) und der ESGRAF-R sowie aufgrund der hohen wissenschaftlichen Expertise der Testentwickler im Bereich der Grammatikdiagnostik bei Kindern ist von einer hohen Inhaltsvalidität des Instrumentes auszugehen.

Eine Facette der Konstruktvalidität überprüft die Zuverlässigkeit der Testergebnisse dadurch, dass der Zusammenhang der Testergebnisse und anderer Variablen untersucht wird (zur theoretischen Fundierung der Konstruktvalidität vgl. Cronbach & Meehl, 1955). In diesem Zusammenhang hat sich die Definition eines sogenannten nomologischen Netzes bewährt, das einen theoretisch angenommenen Zusammenhang bzw. auch Nichtzusammenhänge zwischen den Testvariablen und weiteren Konstrukten bzw. Variablen definiert⁶. Die Annahme eines Zusammenhanges zwischen den Testwerten und

den weiteren Konstrukten wird dabei auch als konvergente Validität bezeichnet, die Annahme eines geringen bzw. eines nicht vorhandenen Zusammenhanges als divergente Validität.

Im Rahmen der Konstruktvalidierung der ESGRAF 4-9 wurde auf Basis theoretischer Vorüberlegungen das in Tabelle 5 dargestellte nomologische Netz definiert, wobei hier besonders auf die Spalte „K-ABC“ hinzuweisen ist, da dieser Untertest als Maß für Intelligenz unter dem Aspekt der divergenten Validität mitberücksichtigt wurde.

Die korrelativen Befunde der Konstruktvalidierung (Produkt-Moment-Korrelation nach Pearson) sind in Tabelle 6 dargestellt.

Wie deutlich wird, lassen sich die meisten im nomologischen Netzwerk aufgestellten Hypothesen über das Vorhandensein von signifikanten bzw. nicht signifikanten Zusammenhängen empirisch bestätigen. Lediglich bei Subtest 2 findet sich ein den Hypothesen nicht entsprechendes Ergebnismuster.

Ein weiterer Aspekt der Konstruktvalidierung bezieht sich auf die Interkorrelationen der Konstrukte bzw. der Subtests der ESGRAF 4-9. Hier zeigt sich auf Ebene der Subtests eine korrelative Struktur, die die qualitative Unterschiedlichkeit der Dimensionen aufzeigt.

Einfluss weiterer Außenkriterien auf die Testergebnisse

Abschließend wurden die Zusammenhänge zwischen den Testwerten, die bei Anwendung der ESGRAF 4-9 resultieren, und Außenkriterien außerhalb der klassischen Konstruktvalidierung untersucht⁷. Hierbei ging es vor allem um das Ziel, die Variablen zu er-

⁶ Der Grundgedanke des nomologischen Netzes ist für jede Form der empirischen Validierung sinnvoll, da der Forscher hier (sinnvollerweise) gezwungen ist, Zusammenhänge (oder Nichtzusammenhänge) zwischen dem von ihm untersuchten Konstrukt auf theoretischer Ebene zu formulieren und dann die entsprechende Erhebung der Daten operational umzusetzen. Im Gegensatz hierzu steht die häufig zu beobachtende Praxis, „alles mit allem“ zu korrelieren, die von Cronbach und Meehl als schwache Form der Konstruktvalidierung bezeichnet wird. Anastasi (1986) bezeichnet diese Form der Konstruktvalidierung als blinden Empirismus.

⁷ Inwieweit diese Untersuchung mit der Ermittlung von Kriteriumsvalidität gleichgesetzt werden kann, bleibt dem Leser überlassen (vgl. zu dieser Diskussion Hartig, Frey & Jude, 2012, S. 162-166). Die Autoren haben sich für die neutralere Bezeichnung „Außenkriterien“ entschieden.

Tabelle 5: Nomologisches Netz als Grundlage für die Konstruktvalidierung der ESGRAF 4-9

	ETS 4-8	SET 5-10	IDIS	K-ABC
Frühe Fähigkeiten				
Subtest 1a Morphologie: Subjekt-Verb-Kongruenz	0	X	0	0
Subtest 1b Syntax: Verbzweitstellungsregel	0	0	X	0
Subtest 3 Genus (Lexikon)	X	0	X	0
Späte Fähigkeiten				
Subtest 2 Syntax: Verbendstellungsregel	0	X	0	0
Subtest 4 Morphologie: Kasus Akkusativ + Dativ	X	X	X	0
Subtest 5 Morphologie: Plural		X	0	X
Subtest 6 Syntax: Passiv	X	X	0	0
Subtest 6 Morphologie: Kasus: Genitiv	X	0	X	0
Subtest 6 Konjunktionalsätze	0	X	0	0
Subtest 6 Gesamtscore	X	X	X	0
Kennwerte				
Mittelwert (Median)	36.39 (37.00)	9.53 (10.50)	36.76 (38.00)	11.06 (11.50)
Standardabweichung	3.739	2.474	7.927	3.953
N	110	110	111	112

Anmerkungen.

ETS 4-8 (Angermaier, 2008): Untertest GE (Grammatische Entwicklung). Der Untertest GE überprüft bei Kindern (4;0-8;11) einen Mix vorwiegend morphologischer Fähigkeiten (Plural, Komparativ, Vergangenheit).

SET 5-10 (Petermann, 2010): Untertest 9 (Erkennen / Korrektur inkorrektur Sätze). Der Untertest überprüft bei Kindern (5;0-8;11) syntaktisches und morphologisches Regelwissen bei jüngeren Kindern deklarativ (Erkennen), bei älteren Kindern zusätzlich prozedural (Korrektur).

IDIS (Schöler, 1999): Untertests SM und SU. Die beiden Untertests überprüfen Akkusativ und Dativ sowohl in der Nominal- als auch in der Präpositionalphrase mit je 24 Items.

K-ABC Untertest „Dreiecke“ (vgl. Melcher & Preuß, 2009): Dieser Untertest ist der K-ABC-Skala „Intellektuelle Fähigkeiten“ zugeordnet und hat von allen Subskalen den größten Erklärungswert.

In den Feldern mit „X“ werden statistisch signifikante Korrelationen erwartet. In den Feldern, die eine „0“ enthalten, wird ausdrücklich von keinem (signifikanten) Zusammenhang ausgegangen.

Tabelle 6: Ergebnisse der Konstruktvalidierung

	ETS 4-8	SET 5-10	IDIS	K-ABC
Frühe Fähigkeiten				
Subtest 1a Morphologie: Subjekt-Verb-Kongruenz		0.212*	.245*	0
Subtest 1b Syntax: Verbzweitstellungsregel		0	.236*	0
Subtest 3 Genus (Lexikon)	.510**	0	.494**	0
Späte Fähigkeiten				
Subtest 2 Syntax: Verbendstellungsregel	.399**	0 (n.s.)	.482**	0
Subtest 4a Morphologie: Kasus Dativ	.551**	.350**	.801**	0
Subtest 4b Morphologie: Kasus Akkusativ	.373**	.319**	.396**	0
Subtest 5 Morphologie: Plural	.278**	0	.302**	0
Subtest 6 Syntax: Passiv	.530**	.320**	0	0
Subtest 6 Morphologie: Kasus: Genitiv	.592**	0	.406**	0
Subtest 6 Konjunktionalsätze	0	.286**	.294**	0
Subtest 6 Gesamtscore	.534**	.289**	.477**	0

Anmerkungen. N = 113, $\alpha < 5\% = *$, $\alpha < 1\% = **$, **fett** markiert sind die erwartungskonformen Korrelationen (entweder signifikant oder mit „0“ bezeichnet nichtsignifikant), *kursiv* sind die Zellen, in denen entweder trotz Erwartung keines Zusammenhanges eine Korrelation vorlag oder trotz Erwartung eines Zusammenhanges keine Korrelation gefunden wurde (hier zu besseren Sichtbarkeit noch der Zusatz n.s. für „nicht signifikant“ ergänzt).

Tabelle 7: Interkorrelation der Subtests/Dimensionen der ESGRAF 4-9

		Subtest							
		1a	1b	2	3	4a	4b	5	6
Subtest	1a	1							
	1b	.929**	1						
	2	.359**	.326**	1					
	3	.226*	.237*	.414**	1				
	4a	.153	.167	.362**	.496**	1			
	4b	.228*	.289**	.390**	.365**	.380**	1		
	5	.144	.148	.358**	.338**	.237*	.274**	1	
	6	.174	.186*	.408**	.381**	.470**	.361**	.271**	1

Anmerkungen. N = 113, $\alpha < 5\% = *$, $\alpha < 1\% = **$, **fett** markiert sind alle signifikanten Korrelationen

mitteln, die einen Einfluss auf die Testergebnisse haben und die unter Umständen Kriterien für eine Normierung darstellen.

Variablen, deren Einfluss unter der Fragestellung einer etwaigen Berücksichtigung bei einer Testnormierung zu untersuchen ist, sind

- das Alter,
- das Geschlecht,
- die Bildung der Eltern sowie
- Umfeld (städtisch, ländlich/suburban).

Weitere Variablen, die zusätzlich erhoben worden sind, waren

- die Einschätzung der grammatischen Fähigkeiten durch Erzieherinnen und Erzieher bzw. Lehrkräfte und
- das Vorliegen einer Sprachstörung.

Die Ergebnisse der Korrelation der Testvariablen⁸ mit den sechs Außenkriterien sind in Tabelle 8 dargestellt.

Tabelle 8: Korrelationen mit den Außenkriterien

	Bildung Eltern	Alter	Stadt/Land	Geschlecht	Einschätzung Lehrer	Sprachauf-fälligkeit
Frühe Fähigkeiten						
Subtest 1 Morphologie: Subjekt-Verb-Kongruenz	-.070	.089	.005	.185	-.087	.206*
Subtest 1 Syntax: Verbzweitstellungsregel	-.062	.064	.004	.087	-.044	.168
Subtest 3 Genus (Lexikon)	-.091	.513**	-.218*	.111	-.018	.010
Späte Fähigkeiten						
Subtest 2 Syntax: Verbendstellungsregel	-.141	.420**	.039	.246**	-.185	-.030
Subtest 4 Morphologie: Kasus Dativ	.014	.407**	-.217*	.221*	-.188*	.066
Subtest 4 Morphologie: Kasus Akkusativ	.076	.317**	-.099	.105	-.115	.010
Subtest 5 Morphologie: Plural	-.098	.297**	.012	.102	-.091	-.068
Subtest 6 Syntax: Passiv	-.191*	.732**	.005	.056	.020	.142
Subtest 6 Morphologie: Kasus: Genitiv	.144	.044	.152	.082	-.153	.008
Subtest 6 Konjunktionalsätze	-.026	.519**	.059	.093	-.112	.126
Subtest 6 Gesamtscore	-.052	.487**	-.113	.068	-.124	.189*

Anmerkungen. N = 113, $\alpha < 5\% = *$, $\alpha < 1\% = **$, **fett** markiert sind alle signifikanten Korrelationen

Die Ergebnisse zeigen eindeutig, dass das Alter der untersuchten Kinder einen Einfluss auf die Testergebnisse hat. Der Einfluss der anderen Variablen ist nur punktuell (wenn auch wie der Zusammenhang zwischen der Bildung der Mutter und der Verwendung des Genitivs) und nicht so durchgängig wie beim Alter.

Diskussion

Die Ergebnisse der testtheoretischen Absicherung belegen, dass der ESGRAF 4-9 den Testgütekriterien Objektivität, Reliabilität und Validität genügt. Das einzige abweichende Ergebnis im Bereich der Reliabilität bei dem fünften Untertest (Plural) war spracherwerbstheoretisch erwartungskonform, da, wie bereits erwähnt, die Pluralmarkierungen im Gegensatz zu anderen grammatischen Strukturen nicht regelgeleitet, sondern eher im 1:1 Lernen von Kindern gelernt werden, so dass keine innere Konsistenz zwischen den Items zu erwarten ist. Dies unterstreicht auch die geringe Übereinstimmung dieses Untertests mit dem entsprechenden Pluralsubtest der ETS 4-8 (siehe Tabelle 6). Da als Ergebnis der aktuell laufenden Normierungsstudie Normdaten für jeden Subtest generiert werden, kann gegebenenfalls zukünftig auf die Durchführung der Pluralüberprüfung verzichtet werden. Bezüglich aller regelhaft erworbenen grammatischen Strukturen liefert der ESGRAF 4-9 hingegen Normdaten, welche unter Berücksichtigung strenger psychometrischer Qualitätsstandards erhoben wurden.

Besonders hervorzuheben sind die Ergebnisse der Konstruktvalidierung, die aufzeigen, dass das a priori formulierte nomologische Netze der empirischen Überprüfung standhält.

Inwieweit bei der anstehenden Testnormierung noch weitere Kriterien außerhalb

des Alters herangezogen werden sollten, muss diskutiert werden. Bei der Überprüfung der Außenkriterien überrascht der Befund, dass weder das Geschlecht der Kinder noch die Bildung der Eltern einen nennenswerten Einfluss auf die Befunde haben. Die geringe Korrelation mit dem Urteil von Erzieherinnen und Erziehern bzw. Lehrkräften kann nach Auffassung durch die Autoren durch die hohe Komplexität und nicht immer direkte Zugänglichkeit des Bewertungsgegenstandes erklärt werden.

Literaturverzeichnis

- Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention*. Heidelberg: Springer.
- Anastasi, A. (1986). Evolving concepts of test validation. *Annual review of Psychology*, 37, 1–15.
- Angermaier, M. (2007). *ETS 4-8*. Frankfurt a. M.: Harcourt Test Services.
- Asendorpf, J. & Wallbott, H.G. (1979). Maße der Beobachterübereinstimmung: Ein systematischer Vergleich. *Zeitschrift für Sozialpsychologie*, 10, 243-252.
- Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler*. Berlin, Heidelberg: Springer.
- Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Esser, G., Wyschkon, A., Ballaschk, K. & Hänsch, S. (2010). *P-ITPA: Potsdam-Illinois Test für Psycholinguistische Fähigkeiten*. Göttingen: Hogrefe.
- Hartig, J., Fray, A. & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.),

⁸ Für die dichotomen Variablen Stadt/Land, Geschlecht und Sprachauffälligkeit wurde eine punktbiseriale Korrelation berechnet, die dem t-Test für unabhängige Stichproben (mit zweifach gestufter unabhängiger Variablen) entspricht (vgl. Bortz & Schuster, 2010, S. 171-172). Rasch, Friese, Hofmann und Naumann (2010, S. 140) weisen nach, dass die Berechnung der punktbiserialen Korrelation der der Produkt-Moment-Korrelation mit einer dichotomen und einer intervallskalierten Variablen entspricht.

- Testtheorie und Fragebogenkonstruktion* (S. 143-171). Berlin, Heidelberg: Springer.
- Kauschke, C. & Siegmüller, J. (2010). *Pathologische Diagnostik bei Sprachentwicklungsstörungen*. München: Urban & Fischer.
- Lees, J. & Urwin S. (1998). *Children with language disorders*. London: Whurr Publishers.
- Lienert, G.A. & Raatz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: Beltz.
- McGraw, K.O. & Wong, S.P. (1996a). Forming Inferences About Some Intraclass Correlation Coefficients. *Psychological Methods*, 1, 30-46.
- McGraw, K.O. & Wong, S.P. (1996b). Correction to McGraw and Wong (1996). *Psychological Methods*, 1, 390.
- Melchers, P. & Preuß, U. (Hrsg.) (2009). *K-ABC: Kaufman – Assessment Battery for Children*. Frankfurt: Pearson.
- Moosbrugger, H. (2012a). Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 227-274). Berlin, Heidelberg: Springer.
- Moosbrugger, H. (2012b). Klassische Testtheorie (KTT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 103-118). Berlin, Heidelberg: Springer.
- Moosbrugger, H. & Kelava, A. (2012a). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7-26). Berlin, Heidelberg: Springer.
- Moosbrugger, H. & Kelava, A. (Hrsg.) (2012b). *Testtheorie und Fragebogenkonstruktion*. Berlin, Heidelberg: Springer.
- Motsch, H.-J. (2000). *ESGRAF: Evozierte Sprachdiagnostik grammatischer Fähigkeiten*. München: E. Reinhardt.
- Motsch, H.-J. (2010). *Kontextoptimierung. Evidenzbasierte Intervention bei grammatischen Störungen in Therapie und Unterricht*. München: E. Reinhardt.
- Motsch, H.-J. (2013). *ESGRAF-R. Modularisierte Diagnostik grammatischer Störungen*. München: E. Reinhardt.
- Motsch, H.-J. & Becker, L.-M. (2014). Grammatikerwerb deutschsprachiger Kinder zwischen 4 und 9 Jahren (GED 4 –9). In *Vierteljahresschrift für Heilpädagogik (VHN)* 1, 71-73.
- Petermann, F. (2010). *SET 5-10. Sprachstandserhebungstest für Kinder im Alter zwischen 5 und 10 Jahren*. Göttingen: Hogrefe.
- Rasch, B., Frieese, M., Hofmann, W. & Naumann, E. (2010). *Quantitative Methoden. Band 1. Einführung in die Statistik für Psychologen und Sozialwissenschaftler*. Heidelberg: Springer.
- Schermelleh-Engel, K. & Werner, C.S. (2012). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 119-141). Berlin, Heidelberg: Springer.
- Schöler, H. (1999). *IDIS-Inventar diagnostischer Informationen bei Sprachentwicklungsauffälligkeiten*. Heidelberg: Schindele.
- Szagan, G. (2007). Grammatikentwicklung. In H. Schöler & A. Welling (Hrsg.), *Sonderpädagogik der Sprache* (S. 29-41). Göttingen: Hogrefe.
- Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

Prof. Dr. Christian Rietz

Universität zu Köln
Fragenheimstr. 4
50931 Köln
Christian.Rietz@uni-koeln.de

Prof. Dr. Hans-Joachim Motsch

Universität zu Köln
Klosterstr. 79b
50931 Köln
j.motsch@uni-koeln.de

Erstmalig eingereicht: 24.04.2014

Überarbeitung eingereicht: 27.05.2014

Angenommen: 10.06.2014