2025, 2. Jg., Heft 2
ISSN 2944-4969
https://doi.org/10.26204/KLUEDO/9339

Stefan Wahl, Iris Rautenberg, Alicia Hückmann & Vanessa Siegel

Messinstrumente zur Erfassung der Groß-/Kleinschreibleistung

Die satzinterne Großschreibung zählt zu den fehleranfälligsten Bereichen der deutschen Orthographie. In der Deutschdidaktik gibt es seit den 90er-Jahren eine intensive Diskussion darüber, wie die Groß-/Kleinschreibung (GKS) in der Schule vermittelt werden sollte. In Evaluationsstudien zur Überprüfung der Effektivität verschiedener didaktischer Ansätze zur GKS stellt die GKS-Leistung der Schüler:innen das zentrale Kriterium dar. In diesem Beitrag wird die GKS-Leistung hinsichtlich der unterschiedlichen Fälle der GKS und des Nutzungskontexts beschrieben. Dazu werden drei Messverfahren vorgestellt, die in der Evaluationsstudie KeGS zum Einsatz kamen und unterschiedliche Bereiche der GKS-Leistung erfassen: ein Test zur Erfassung der Leistung bei der Produktion isolierter GKS-Entscheidungen (PRO-ISO), ein Test zur Erfassung der Leistung bei der Produktion integrierter GKS-Entscheidungen (PRO-INT) und ein Test zur Erfassung der rezeptiven GKS-Leistung beim syntaktischen Lesen (REZ-INT). Anhand der Messdaten der Studie, in der die drei Tests 235 Siebtklässler:innen vorgelegt wurden, werden die Messverfahren vorgestellt und deren Zusammenhänge sowie ihre Vor- und Nachteile diskutiert.

Schlagwörter: Groß-/Kleinschreibung, Testkonstruktion, Leistungsmessung, Evaluationsstudie

Measurement Methods for Assessing Capitalization Performance

Sentence internal capitalization is one of the most error-prone areas of German orthography. Since the 1990s, there has been an ongoing debate in German didactics about how capitalization should be taught in schools. In studies evaluating the effectiveness of various didactic approaches, students' capitalization performance serves as the central criterion. This article describes capitalization performance with regard to different cases of capitalization and the mode of use. Three tests, which had previously been used in the evaluation study KeGS, are introduced, each targeting different areas of capitalization performance: a test for productive, isolated capitalization decisions (PRO-ISO), a test for productive, integrated capitalization decisions (PRO-INT), and a test for the receptive capitalization performance during syntactic reading (REZ-INT). Using data from the study in which the three tests were presented to 235 seventh-grade students, this article presents the measurement methods, their interdependence, as well as their advantages and disadvantages.

Keywords: Capitalization, Test Construction, Performance Measurement, Evaluation Study

1 Einleitung

Die besondere Schwierigkeit der Groß- und Kleinschreibung (GKS) ist seit langem hinreichend belegt (z. B. Pießnack & Schübel 2005; Steinig & Betzel 2014; Müller 2016; Fuhrhop & Romstadt 2021). Problematisch ist insbesondere die Großschreibung von vorwiegend abstrakten Substantiven und von Substantivierungen (u. a. Rautenberg & Wahl 2024). Dafür mitursächlich gilt der traditionelle wortartbasierte Vermittlungsansatz, der in der Unterrichtspraxis oftmals dominiert und vor allem die Großschreibung von lexikalischen Substantiven, insbesondere von prototypischen Konkreta, sicher erfasst (Röber-Siekmeyer

1999; Bredel 2006). Fälle, deren Schreibung syntaktisch bedingt ist, müssen im fortgeschrittenen Erwerb als Ausnahmefälle deklariert und mühsam eingeübt werden, insbesondere wenn zuvor verkürzte Kleinschreibregeln (‚Verben und Adjektive schreibt man klein') zum Einsatz kamen. Konsequent syntaktische Vermittlungsansätze knüpfen die Großschreibung hingegen von Beginn an an bestimmte Merkmale großgeschriebener Wörter im Satzkontext und decken damit die meisten Fälle zuverlässig ab. Allerdings erfassen auch syntaxbasierte Ansätze nicht alle Großschreibungen (Gallmann 1997). Als Sonderfälle besprochen werden müssen u. a. Schreibungen, die rein lexikalisch bedingt sind (z. B. *Schlange stehen*) (Günther & Nünke 2005). Da der Peripheriebereich jedoch vergleichsweise überschaubar ist, stellen syntaktische Ansätze eine vielversprechende Alternative dar (u. a. Gaebert 2012; Wahl et al. 2017). Sie zeichnen sich durch eine linguistische Fundierung (z. B. Maas 1992; Eisenberg 2020) aus und verbinden die GKS mit grundlegendem grammatischem Wissen, da sie ganze Phrasen und Sätze in den Blick nehmen. In einem syntaxbasierten Unterricht wird die Aufmerksamkeit gezielt auf die Erweiterbarkeit nominaler Kerne durch Adjektivattribute (Röber-Siekmeyer 1999), die Flexion der Nominalphrase (NP) (Funke 1995) oder syntaktische Kontraste (Funke et al. 2013) gelenkt.

In empirischen Studien zur Evaluation der didaktischen Ansätze sind die GKS-Leistungen der Schüler:innen (SuS) das zentrale Bewertungskriterium für deren Wirksamkeit. Im Rahmen der Interventionsstudie *KeGS*[1] (Kompetenzentwicklung Großschreibung in der Sekundarstufe), in der die Effektivität von drei didaktischen Ansätzen zur GKS miteinander verglichen wurde (Rautenberg et al. 2025; Hückmann et al. eingereicht), wurden drei verschiedene Messverfahren entwickelt, die zentrale Aspekte der GKS-Leistung erfassen.

2 Fragestellungen

Die Fragestellungen dieses Beitrags beziehen sich zum einen auf pragmatische, forschungsmethodische Aspekte, die sich direkt aus dem Projektkontext der Evaluationsstudie ergaben: Auf welche Weise kann in einer empirischen Evaluationsstudie die GKS-Leistung gemessen werden, sodass die Messinstrumente einerseits praktikabel und mit einem zumutbaren Zeitaufwand in großen Stichproben in der Schule eingesetzt werden können und dass sie andererseits möglichst differenzierte Informationen zu wichtigen Nutzungskontexten der GKS-Leistung liefern? Wie sind die Objektivität, Reliabilität, Validität und Praktikabilität dieser Tests ausgeprägt? Zum anderen liegen dem Beitrag theoretisch-konzeptuelle Aspekte zugrunde: In welchem Zusammenhang stehen die unterschiedlichen Nutzungskontexte und Teilkompetenzen der GKS-Leistung? Welche Interpretationen erlauben die durch die Tests erfassten Bereiche der GKS-Leistung und welche Limitationen sind gegeben?

[1] Das Projekt wurde durch die DFG gefördert (Projektnummer: 451577334).

3 Messinstrumente in Evaluationsstudien zur Groß-/Kleinschreibung

In bisherigen Studien zur Evaluation didaktischer Ansätze kommen unterschiedliche Messinstrumente zum Einsatz. Echtwortlückendiktate sind das häufigste Erhebungsformat und wurden in den folgenden Studien verwendet:

- Betzel (2015): 40 Sätze mit 53 Großschreibfällen (Konkreta, Abstrakta, +/- abgeleitete Substantive, +/- lexikalisierte Substantivierungen, NPs mit/ohne Artikel/Adjektiv) und 50 Kleinschreibfällen
- Funke et al. (2013): zehn Sätze mit 20 genuinen Substantiven
- Wahl et al. (2017): 15 Sätze mit 16 Großschreibfällen (Konkreta, Abstrakta, substantivierte Adjektive/Verben, NPs mit/ohne Artikel/Adjektiv) und zwölf Füllwörtern
- Luxemburger Studien (Brucher et al. 2020; Mangelschots et al. 2023; Weth et al. 2024): 36–48 Zielitems (Konkreta, Abstrakta, Substantivierungen, NPs mit/ohne Artikel/Adjektiv) und 24 Füllwörter

Seltener werden Tests verwendet, bei denen SuS in vollständig vorgegebenen Sätzen über die GKS entscheiden müssen (z. B. Wahl et al. 2017; Bangel 2022; Mangelschots et al. 2023; Weth et al. 2024).[2] Bei Wahl et al. (2017) werden den Studienteilnehmenden insgesamt 96 Testsätze mit je einem Testwort vorgegeben, die in zwölf Testversionen à acht Testsätzen präsentiert werden, wobei die verschiedenen semantisch-lexikalischen und syntaktischen Merkmale systematisch variiert sind. Die anderen Studien nutzen Tests mit 20 Sätzen (Bangel 2022) bzw. sechs bis acht Sätzen (Mangelschots et al. 2023; Weth et al. 2024). Die Tests von Bangel (2022) und Mangelschots et al. (2023) unterscheiden sich insofern von den anderen beiden, als die Sätze weitgehend in Kleinschreibung vorgegeben werden und für jedes Wort eine Entscheidung getroffen werden muss. Bei Wahl et al. (2017) und Weth et al. (2024) müssen die SuS die passende von zwei vorgegebenen Schreibalternativen auswählen.

Bei Funke et al. (2013) wird ein Test zum syntaktischen Lesen eingesetzt. Der Test enthält acht Aufgaben, in denen SuS die GKS von je einem lexikalischen Verb in einem Satzanfang korrekt interpretieren müssen, um das passende Satzende zu finden. Ein differenzierteres Testverfahren, bei denen die SuS je zwei Verständnisaufgaben zu zwölf vollständig vorgegebenen Kurztexten bearbeiten müssen (allerdings nicht in Verbindung mit einer Intervention), findet sich in Funke und Sieger (2009). Die Kurztexte enthalten einen strukturell ambigen Satz, der durch die GKS disambiguiert wird. In den Aufgaben werden die SuS dazu aufgefordert, den Satz zunächst in eigenen Worten zu paraphrasieren, indem sie eigenständig einen vorgegebenen Satzanfang vervollständigen. Anschließend beantworten sie eine Single-Choice Aufgabe. Anhand der Antworten wird deutlich, ob die SuS die GKS beim Lesen beachten und die elizitierte syntaktische Information über mehrere Bearbeitungsschritte hinweg verfügbar halten können.

[2] Es werden nur Testverfahren berücksichtigt, die in Evaluationsstudien zum Einsatz kamen. Vergleichbare Messinstrumente wurden darüber hinaus auch in anderen Studien, u. a. zur Ermittlung von Itemschwierigkeiten (Müller 2016), eingesetzt.

4 Aspekte der Groß-/Kleinschreibleistung in der Evaluationsstudie *KeGS*

Welches Kompetenzziel soll durch den GKS-Unterricht in der Schule erreicht werden? Was ist damit gemeint, wenn man SuS zuschreibt, dass sie die GKS (nicht) (gut) beherrschen? Zur Durchführung der Evaluationsstudie *KeGS* wurde zunächst inhaltlich genauer bestimmt, welche Aspekte der GKS-Leistung erfasst werden sollen, um dafür im Anschluss geeignete Messverfahren entwickeln und einsetzen zu können. In forschungsmethodischem Vokabular ausgedrückt: Für die abhängige Variable (aV) mussten passende Operationalisierungen geschaffen werden.

4.1 Nutzungskontexte der Groß-/Kleinschreibung

Es wurden drei Testverfahren entwickelt, die jeweils einen anderen wichtigen Nutzungskontext der GKS erfassen. Der erste Kontext ist die GKS beim Schreiben, was sowohl das freie Schreiben in offenen Schreibsituationen (z. B. Verfassen einer schriftlichen Erzählung) als auch das Schreiben vorgegebener Sätze/Texte (z. B. bei einem Diktat) bedeuten kann. Das für diesen Kontext entwickelte Messinstrument bezieht sich aus methodischen Gründen (s. u.) auf das Schreiben vorgegebener Sätze in einem Lückendiktat. Nach Bredel (2013, 105) ist für die Bewältigung dieser Aufgabe implizites (primärsprachliches) Wissen erforderlich. Der zweite Nutzungskontext ist das Treffen von nur auf den Rechtschreibaspekt der GKS bezogenen Entscheidungen, z. B. wenn die GKS in eigenen oder fremden Sätzen/Texten korrigiert werden soll. Diese GKS-Entscheidungen bezeichnen wir als isoliert; im Gegensatz zum ersten Kontext, bei dem die GKS-Entscheidungen in den Prozess der vollständigen Wortschreibung integriert sind und daher nicht allein im Fokus der Aufmerksamkeit stehen. Erforderlich ist für die erfolgreiche Bearbeitung dieser Aufgabe nach Bredel metasprachliches Wissen, das u. a. mit einer Deautomatisierung der Prozesse einhergeht. Der dritte Nutzungskontext bezieht sich auf die Wahrnehmung der durch die GKS markierten Information (z. B., dass ein großgeschriebenes Wort der Kern einer NP ist) und deren semantische Interpretation für das sinnerfassende Lesen. Nach Bredels Wissenstaxonomie (ebd.) wäre auch hier metasprachliches Wissen (in Form von integriertem Prozesswissen) erforderlich. Die ersten beiden Nutzungskontexte der GKS bezeichnen wir als produktiv[3] (Schreiben), den dritten als rezeptiv (Lesen).

4.2 Berücksichtigte Fälle der Groß-/Kleinschreibung

Großgeschrieben wird laut Amtlichem Regelwerk (2024) bei „Überschriften, Werktiteln und dergleichen, Satzanfängen, Substantiven und Substantivierungen, Eigennamen mit ihren nichtsubstantivischen Bestandteilen, bestimmten festen nominalen Wortgruppen mit nichtsubstantivischen Bestandteilen, Anredepronomen und Anreden“. Substantive und Substantivierungen (im Folgenden als *satzinterne Großschreibung* bezeichnet) stellen die größte Fehlerquelle dar, wobei es Schwierigkeitsunterschiede gibt (vgl. u. a. Rautenberg & Wahl 2024; Bangel 2022; Betzel 2015).

[3] Eingewendet werden kann allerdings, dass auch bei der Entscheidung über die GKS bei vorgelegten Sätzen rezeptive Fähigkeiten eine Rolle spielen (da die Sätze gelesen und syntaktisch verarbeitet werden müssen) und diese Aufgabe daher nicht rein *produktiv* ist.

Tab. 1: Schwierigkeitsunterschiede bei der GKS

Schwierigkeitsunterschiede bei der GKS (schwieriger > einfacher)	
Großschreibung	Substantivierungen (Adjektiv > Verb) > Abstrakta > Konkreta
	Kern in NP isoliert > Kern in NP expandiert (mit Artikel und/oder Attribut)
	ohne Suffix > mit Suffix
Kleinschreibung	Denominalisierungen > andere Fälle
	Adjektivattribut nach Artikelwort > andere Fälle
	(dazu eine Diskussion in Rautenberg & Wahl 2024)

Die Einflussfaktoren (vgl. Tab 1) sind nicht disjunkt. In jedem einzelnen Fall handelt es sich um ein Konkretum, Abstraktum oder eine Substantivierung, das bzw. die unabhängig davon eine bestimmte morphologische Komplexität hat und wiederum unabhängig davon in einen bestimmten syntaktischen Kontext eingebettet ist. Bei der Konstruktion der Tests wurden die Merkmale daher – soweit möglich – systematisch kombiniert, sodass die Messungen eines Merkmals immer mit dem ganzen Spektrum der Ausprägungen der anderen Merkmale auftreten und deren Einfluss dadurch „kontrolliert" bzw. konstant gehalten ist.

5 Entwicklung von Messverfahren zur Groß-/Kleinschreibleistung

In der Evaluationsstudie *KeGS* (Rautenberg et al. 2025; Hückmann et al. eingereicht) wurden drei Messinstrumente zur Erfassung der GKS-Leistung konstruiert, um differenzierte Aussagen über verschiedene Fälle der GKS in unterschiedlichen Nutzungskontexten machen zu können. Sie wurden direkt vor der Intervention (Prätestung), direkt nach der Intervention (Posttestung) und etwa drei Monate später (Follow-up) eingesetzt. Bei der Analyse werden nur die Prätest-Daten berücksichtigt, weil die Leistungen zu diesem Zeitpunkt noch nicht durch die verschiedenen Interventionen beeinflusst sind und daher die Daten aller experimentellen Gruppen zusammengefasst werden können.

5.1 Stichprobe

An der Studie nahmen 247 Siebtklässler:innen aus Baden-Württemberg teil. Im Gegensatz zu früheren Interventionsstudien (u. a. Wahl et al. 2017; Brucher et al. 2020; Bangel 2022) berücksichtigt die Studie damit ältere Lerner:innen. Die Daten von zwölf SuS mit einer LRS-Diagnose wurden nicht berücksichtigt. Von den 235 SuS, deren Daten in die Analyse eingingen, besuchten 37 (15,7 %) eine Werkrealschule, 36 (15,3 %) eine Gemeinschaftsschule, 129 (54,9 %) eine Realschule und 33 (14,0 %) ein Gymnasium.

5.2 Auswertungsverfahren

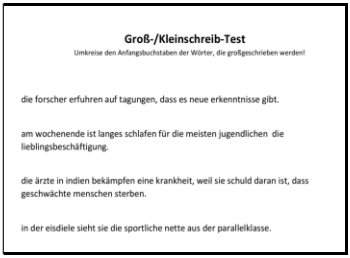
Zuerst wurde bei allen drei Tests und für jeden darin gemessenen Aspekt eine Rasch-Skalierung durchgeführt. Diese Skalenanalysen haben im Gegensatz zur sog. Klassischen Testtheorie den Vorteil, dass – sofern Modellpassung vorliegt – die Itemparameter unabhängig von der spezifischen Stichprobe und die Personenparameter unabhängig von der gewählten Itemmenge geschätzt werden können (Hambleton et al. 1991). Dadurch können auch Ergebnisse aus verschiedenen Tests auf einer gemeinsamen Schwierigkeitsskala miteinander verglichen werden, was es ermöglichte, die Ergebnisse der drei Testverfahren in Beziehung zueinander zu setzen.

Schwierigkeitsunterschiede zwischen Aspekten wurden durch Varianzanalysen auf statistische Signifikanz überprüft. Bei Verletzungen der Sphärizitäts-Annahme (getestet mit dem Mauchly-Test) wurden Korrekturen nach Greenhouse-Geisser durchgeführt. Der Zusammenhang zwischen Aspekten wurde durch Korrelationsanalysen nach Pearson überprüft.

5.3 Test zur Erfassung der Leistung bei der Produktion isolierter GKS-Entscheidungen (PRO-ISO)

Mit dem Messinstrument PRO-ISO wurde getestet, wie gut die SuS entscheiden können, ob Wörter groß- oder kleingeschrieben werden, wenn ihr Fokus ausschließlich auf diese Entscheidung gerichtet ist und ihre Aufmerksamkeit sowie andere Verarbeitungsressourcen nicht durch andere Anforderungen des Schreibens (z. B. die Wortschreibung) beansprucht werden. Es geht hier also um einen produktiven Aspekt der GKS, der isoliert von anderen Prozessen der Rechtschreibung angesprochen und erhoben wird.

Testkonstruktion. Der PRO-ISO besteht aus 16 Sätzen, die in durchgängiger Kleinschreibung vorgegeben werden (vgl. Abb. 1 links). Die Aufgabe der SuS ist, diejenigen Anfangsbuchstaben zu umkreisen, die großgeschrieben werden. Sie benötigten dafür zwischen 10 und 25 Minuten.



Groß-/Kleinschreib-Test

Umkreise den Anfangsbuchstaben der Wörter, die großgeschrieben werden!

die forscher erfuhren auf tagungen, dass es neue erkenntnisse gibt.

am wochenende ist langes schlafen für die meisten jugendlichen die lieblingsbeschäftigung.

die ärzte in indien bekämpfen eine krankheit, weil sie schuld daran ist, dass geschwächte menschen sterben.

in der eisdiele sieht sie die sportliche nette aus der parallelklasse.

Großschreibungen	Satzanfänge 16	Namen/ Titel 3	Konkreta 8 (+ 10)	
	Abstrakta		*Nominalisierungen*	
	ohne Suffix	*mit Suffix*	*von Verben*	*von Adjektiven*
NP > N	1 (+ 1)	1	1	1
NP > Det N	1 (+ 6)	1 (+ 2)	1 (+ 1)	1
NP > A N	1 (+ 1)	1 (+ 1)	1	1
NP > Det A N	1 (+ 1)	1	1	1
Kleinschreibungen	NP > Det A N 4 (+ 2)		Sonstige 0 (+ 121)	

Abb. 1: Beispielitems und Systematik der Testkonstruktion des PRO-ISO

Die Konstruktion der Testsätze folgte einer Systematik, mit der relevante Fälle der GKS abgedeckt werden (vgl. Abb. 1 rechts). Der Test sollte vier Abstrakta ohne Suffix, vier Abstrakta mit Suffix, vier Nominalisierungen von Verben und vier Nominalisierungen von Adjektiven enthalten. Jeweils eines dieser Testwörter sollte in unterschiedlichen syntaktischen Kontexten auftreten: in NPs ohne Artikel und ohne Attribut (NP > N), mit Artikel und ohne Attribut (NP > Det N), ohne Artikel und mit Attribut (NP > A N) und mit Artikel und mit Attribut (NP > Det A N). Zu jedem dieser 16 Testwörter wurde ein Satz konstruiert, sodass auch 16 Satzanfänge großzuschreiben waren. In den vier Testsätzen, in denen eine NP mit Artikel (in Distanzstellung) und Adjektivattribut vorkam (NP > Det A N), konnte auch die Kleinschreibung der Attribute getestet werden. Die 16 Sätze wurden weiterhin so formuliert, dass auch die anderen Fälle vorkommen. Es sollten drei Namen oder Titel, acht Konkreta und zwei Denominalisierungen enthalten sein. Bei der Ausformulierung der

Sätze haben sich noch weitere Beispiele für die jeweiligen Fälle ergeben, die bei der Auswertung berücksichtigt werden konnten (die Häufigkeiten sind in Abb. 1 rechts jeweils in Klammern angegeben). Die 16 Sätze enthalten 121 sonstige kleinzuschreibende Wörter.

Ergebnisse. Nahezu alle Items in den verwendeten Skalen haben einen akzeptablen gewichteten Fit zum probabilistischen Testmodell; die standardisierten *MNSQ*-Werte liegen im Bereich T ≤ +/-2 (vgl. Tab. 2). Bei den sonstigen Kleinschreibungen liegt bei sieben der 121 Items keine Modell-Passung vor.

Die EAP/PV-Reliabilitäten der einzelnen Skalen sind in Tab. 2 dargestellt. Sie befinden sich in einem mittleren, teilweise in einem guten Bereich zwischen 0.44 und 0.84. Die EAP/PV-Reliabilität der Skala *Namen/Titel* (EAP/PV = 0.23) ist ungenügend. Die Trennschärfe der Skalen ist sehr hoch, alle Skalen haben eine Trennschärfe von TR ≥ .869. Da der Test nicht für individualdiagnostische Zwecke verwendet, sondern in Evaluationsstudien eingesetzt wird, um die Mittelwerte von Stichproben zu vergleichen, sind die Trennschärfe-Reliabilitäten das relevantere Kriterium und die Reliabilität der Skalen kann als gegeben angenommen werden.

Tab. 2: Statistische Kennwerte zu den Aspekten des PRO-ISO

PRO-ISO	Gewichteter Fit		Reliabilität		Schwierigkeit	
	MNSQ	*T*	*EAP/PV*	*TR*	*M*	*SD*
Großschreibungen						
Konkreta	[0.82 ; 1.11]	[-1.3 ; 1.5]	.714	.961	2.87	1.47
Abstrakta	[0.85 ; 1.18]	[-1.9 ; 1.5]	.840	.985	1.75	1.72
ohne Suffix	[0.90 ; 1.15]	[-1.1 ; 2.0]	.804	.988	1.74	1.73
mit Suffix	[0.87 ; 1.17]	[-1.2 ; 1.7]	.659	.970	1.76	1.41
Nominalisierungen	[0.90 ; 1.10]	[-1.7 ; 1.2]	.636	.979	-0.64	1.29
von Verben	[0.91 ; 1.04]	[-1.5 ; 0.7]	.540	.980	-0.50	1.39
von Adjektiven	[0.91 ; 1.04]	[-1.1 ; 0.5]	.487	.991	-0.73	1.50
N	[0.96 ; 1.09]	[-0.7 ; 1.2]	.483	.990	-0.24	1.33
Det N	[0.91 ; 1.27]	[-0.9 ; 1.4]	.733	.986	1.77	1.43
A N	[0.92 ; 1.13]	[-1.0 ; 1.4]	.681	.990	0.27	1.69
Det A N	[0.87 ; 1.06]	[-0.9 ; 1.1]	.672	.994	0.90	2.07
Anfänge	[0.52 ; 1.32]	[-1.7 ; 1.0]	.458	.869	3.22	1.78
Namen/Titel	[1.00 ; 1.01]	[0.1 ; 0.2]	.223	.993	0.64	1.56
Kleinschreibungen						
Adjektive in Distanzstellung	[0.89 ; 1.08]	[-0.3 ; 1.2]	.526	.976	2.46	1.47
Sonstige Kleinschreibung	[0.57 ; 1.48]	[-3.9 ; 3.0]	.628	.899	3.78	1.18

Die Mittelwerte der im PRO-ISO gemessenen Fähigkeitsparameter in Bezug auf die verschiedenen Fälle der GKS sind in Abb. 2 nach Schwierigkeiten geordnet und in Tab. 2 nach Bereichen geordnet dargestellt. Nominalisierungen (*M* = -0.64) sind am schwierigsten und Abstrakta (*M* = 1.75) sind schwieriger als Konkreta (*M* = 2.87). Die Unterschiede sind signifikant ($F(1.9, 444.5) = 826.5$; $p < .001$; $\eta^2 = .779$). Abstrakta mit Suffix (*M* = 1.76) und ohne Suffix (*M* = 1.74) unterscheiden sich nicht ($F(1, 234) = 0.72$; $p = .789$). Nominalisierungen von Verben (*M* = -0,50) sind leichter als von Adjektiven (*M* = -0.73) ($F(1, 234) = 4.82$; $p = .029$; $\eta^2 = .020$).

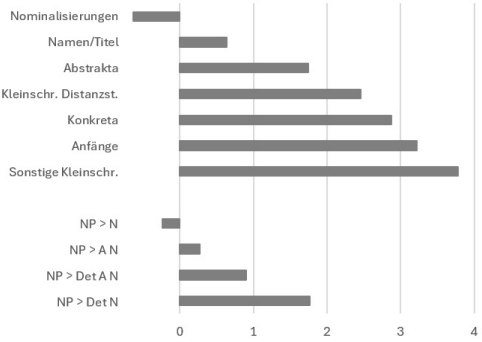


Abb. 2: Mittelwerte der Fähigkeitsparameter im PRO-ISO

Auch die Schwierigkeiten der verschiedenen syntaktischen Kontexte unterscheiden sich deutlich (F(2.7, 641.7) = 161.4; p < .001; η^2 = .408). Steht direkt vor dem nominalen Kern ein Artikel (M = 1.77), ist die Großschreibung am leichtesten, mit Artikel und Adjektivattribut (M = 0.90) schwerer, gefolgt von NPs ohne Artikel, aber mit Attribut (M = 0.27). NPs ohne Artikel und ohne Attribut (M = -0.24) sind am schwierigsten.

Die Kleinschreibung von Adjektiven in NPs mit Distanzstellung (M = 2.46) ist schwieriger als die Kleinschreibung sonstiger Wörter (M = 3.78); dabei wurden am wenigsten Fehler gemacht.

Bewertung. Der PRO-ISO wird auf einem standardisierten Testblatt und mit einer standardisierten Instruktion vorgegeben, sodass von einer hohen Durchführungsobjektivität ausgegangen werden kann. Die EAP/PV-Reliabilitäten befinden sich in einem mittleren, die Trennschärfen in einem sehr guten Bereich (s. o.). In der Evaluationsstudie *KeGS* haben die Interventionen bei allen Skalen des Tests zu signifikanten Verbesserungen vom Prä- zum Posttest geführt (vgl. Rautenberg et al. 2025), was deutlich macht, dass sie über eine hohe Änderungssensitivität verfügen.

Der Test bezieht sich auf den spezifischen Nutzungskontext, dass über die GKS isoliert und unabhängig vom sonstigen Schreibprozess entschieden werden muss. Für Situationen, in denen SuS Schreibfehler in fremden oder eigenen Texten korrigieren, kann bei diesem Test von einer guten ökologischen Validität ausgegangen werden. Inwieweit dieser Test eine valide Messung für die Leistungen in anderen Nutzungskontexten darstellt, kann unter anderem durch dessen Korrelationen mit den Messungen der beiden Verfahren in den anderen Nutzungskontexten beurteilt werden (siehe 5.6). Innerhalb des intendierten Nutzungskontextes ermöglicht der PRO-ISO eine differenzierte Messung der relevantesten Fälle der GKS, sodass er diesbezüglich eine hohe Konstruktvalidität hat.

Die gefundenen Schwierigkeitsunterschiede entsprechen weitgehend den Befunden bisheriger Studien, was auf eine gute Übereinstimmungsvalidität hinweist. Unerwartet ist die hohe Schwierigkeit der Skala *Namen/Titel*. Beim schwierigsten Testwort *Heimische* in der Phrase *die Ausstellung Heimische Wälder* könnten die Fehler durch die Einordnung als

kleinzuschreibendes Attribut erklärt werden. Zur Erklärung des erhöhten Fehleraufkommens bei den Eigennamen *Indien* und *Henry* sind weitere qualitative Analysen notwendig.[4]

Die Ergebnisse zu den Leistungen bei den Kleinschreibungen könnten aufgrund des Aufgabenformats nach oben verzerrt sein. Da die Wörter schon in Kleinschreibung präsentiert wurden, muss bei einem kleinzuschreibenden Wort keine Markierung vorgenommen werden. Aufgrund dessen werden auch Wörter als korrekt bewertet, die gar nicht beachtet oder bei denen keine expliziten Kleinschreibentscheidungen getroffen wurden.

5.4 Test zur Erfassung der Leistung bei der Produktion integrierter GKS-Entscheidungen (PRO-INT)

Mit diesem Test wird die produktive GKS-Leistung getestet, wenn sie integriert in den vollständigen Prozess des Rechtschreibens erbracht werden muss. Die GKS liegt nicht allein im Fokus der SuS, sondern die korrekte Wortschreibung insgesamt. Daher werden ihre Verarbeitungsressourcen auch dadurch beansprucht, über die korrekte Buchstabenfolge eines Wortes zu entscheiden. Ob der erste Buchstabe eines Wortes großgeschrieben wird, ist dabei eine von mehreren zu treffenden Entscheidungen.

Testkonstruktion. Der PRO-INT ist ein Lückendiktat bestehend aus vier Sätzen und 48 zu schreibenden Wörtern. Zu Beginn werden die SuS darauf hingewiesen, dass sie auf die korrekte Rechtschreibung achten sollen, auf die GKS wird aber nicht gesondert hingewiesen. Beim Diktieren wird immer zuerst der gesamte Satz vorgelesen, danach langsam zum Mitschreiben die einzelnen Wörter und zum Schluss nochmals der ganze Satz, um ihnen eine Selbstkontrolle oder das Ergänzen fehlender Teile zu ermöglichen. Die SuS schreiben auf einen Testbogen, bei dem einige Wörter schon vorgegeben sind (vgl. Abb. 3 links). Die Wörter werden auf Linien geschrieben, deren Länge etwa proportional zur Wortlänge ist. Diese Vorgaben sollen den SuS die Orientierung erleichtern und die für den Test benötigte Durchführungszeit etwas reduzieren.

Großschreibungen	Namen/ Titel			
	3			
	Konkreta	*Abstrakta*		*Nominalisierungen*
		ohne Suffix	*mit Suffix*	*von Verben*	*von Adjektiven*
NP > N	*1*		1		1
NP > Det N	*1*		1		1
NP > A N	*1*	1		1	
NP > Det A N	*1*	1		1	

Kleinschreibungen	NP > Det A N	Sonstige
	3	0 (+ 25)

Abb. 3: Testbogen und Systematik der Testkonstruktion des PRO-INT

[4] Mit den Hintergründen der Fehlerschwerpunkte beschäftigt sich auch eine an die *KeGS*-Studie angegliederte Dissertationsstudie (Hückmann, in Vorbereitung).

Bei der Konstruktion der Sätze wurde ebenfalls versucht, möglichst viele Fälle der GKS abzudecken (vgl. Abb. 3 rechts), aber aufgrund des Diktatformats kann die GKS in einer für die SuS angemessenen Bearbeitungszeit nur anhand von deutlich weniger Wörtern getestet werden. Es kamen vier Konkreta, zwei Abstrakta mit und zwei ohne Suffix und zwei verbale und zwei adjektivische Nominalisierungen als Testwörter vor. Die folgenden vier syntaktischen Kontexte werden durch je einen Vertreter der drei oben genannten Kategorien abgedeckt: eine NP mit Artikel oder Adjektivattribut, mit beidem oder mit keinem von beiden. Als Testwörter gibt es zusätzlich drei Namen bzw. Titel, drei Adjektivattribute in NPs mit Distanzstellung und 25 sonstige kleinzuschreibende Wörter. Die Durchführung des Diktats dauerte ca. zehn Minuten.

Ergebnisse. In Tab.3 sind die statistischen Kennwerte zu den Skalen dieses Tests dargestellt. Die standardisierten gewichteten MNSQ-Werte befinden sich bei allen Skalen im Bereich T ≤ +/-2; es liegen also akzeptable Modellpassungen vor.

Die EAP/PV-Reliabilitäten befinden sich im ungenügenden bis mittleren Bereich zwischen 0.23 und 0.66. Die Trennschärfen der Skalen sind im Gegensatz dazu sehr hoch (alle TR ≥ .900). Auch bei diesem Test zeigt sich eine Diskrepanz zwischen (teilweise ungenügenden) Reliabilitäten und (sehr guten) Trennschärfen. Dies wird nach der Darstellung der einzelnen Tests für alle gemeinsam kritisch diskutiert (s. u.).

Tab. 3: Statistische Kennwerte zu den Aspekten des PRO-INT

PRO-INT	Gewichteter Fit		Reliabilität		Schwierigkeit	
	MNSQ	*T*	*EAP/PV*	*TR*	*M*	*SD*
Großschreibungen						
Konkreta	[0.86 ; 1.03]	[-1.1 ; 0.9]	.416	.900	2.15	0.84
Abstrakta	[0.95 ; 1.10]	[-0.3 ; 1.2]	.486	.979	1.68	1.20
Nominalisierungen	[0.97 ; 1.03]	[-0.7 ; 1.1]	.481	.948	-0.48	1.28
N	[0.98 ; 1.01]	[-0.4 ; 1.0]	.230	.990	1.17	1.49
Det N	[0.96 ; 1.01]	[-0.4 ; 1.0]	.280	.996	1.18	1.52
AN	[0.89 ; 1.07]	[-0.8 ; 1.2]	.468	.993	1.36	0.83
Det AN	[0.87 ; 1.08]	[-0.9 ; 1.7]	.506	.993	1.78	1.82
Namen/Titel	[0.81 ; 1.02]	[-2.1 ; 0.2]	.661	.994	1.18	1.52
Kleinschreibungen						
Adjektive in Distanzstellung	[0.99 ; 1.03]	[-0.2 ; 0.3]	.350	.982	1.39	1.12
Sonstige Kleinschr.	[0.94 ; 1.06]	[-0.2 ; 0.8]	.233	1.000	1.18	0.93

Die mittleren Fähigkeitswerte der Skalen des PRO-INT sind in Tab. 3 und in Abb.4 dargestellt.

Auch hier sind Konkreta (M = 2.15) leichter als Abstrakta (M = 1.68) und diese wiederum leichter als Nominalisierungen (M = -0.48). Die Mittelwerte dieser drei Fälle unterscheiden sich signifikant ($F(1.8, 409.7) = 566.2$; $p < .001$; $\eta^2 = .708$).

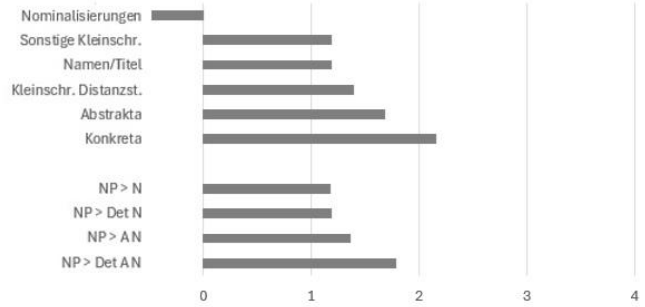


Abb. 4: Mittelwerte der Fähigkeitsparameter im PRO-INT

Auch die syntaktischen Kontexte unterscheiden sich insgesamt signifikant (F(2.8, 665.4) = 12.1; $p < .001$; $\eta^2 = .049$), wobei NPs mit Artikel und Adjektivattribut signifikant leichter als die anderen drei syntaktischen Kontexte sind und die NPs ohne Artikel und mit Adjektivattribut leichter als die beiden Kontexte ohne Attribuierung. Letztere beide Formen unterscheiden sich nicht.

Bewertung. Auch der PRO-INT ist ein relativ objektives Messinstrument, weil der Diktattext und der Erhebungsbogen standardisiert vorgegeben sind. Einschränkungen der Durchführungsobjektivität könnten sich aus Unterschieden beim Diktieren (Geschwindigkeit, Pausen, Betonungen) ergeben. Die EAP/PV-Reliabilitäten der Skalen fallen relativ schlecht aus. Der Anteil der Varianz, der auf Messfehler zurückgeht, ist offenbar verhältnismäßig groß. Dies liegt vermutlich vor allem daran, dass pro Skala maximal drei, teilweise sogar nur zwei Testwörter vorliegen (vgl. Abb. 3). Die geringe Reliabilität des Tests ist damit auch ein pragmatischer Kompromiss zugunsten der erfassten Bandbreite. Die insbesondere für Evaluationsstudien relevanten Reliabilitätsmaße der Trennschärfe (s. o.) und der Änderungssensitivität (vgl. Rautenberg et al. 2025) sind beim PRO-INT sehr gut.

Die Inhaltsvalidität des Tests ist als relativ hoch einzuschätzen, weil der Text trotz seiner Kürze sehr viele Fälle der GKS abdeckt. Auch die ökologische Validität des PRO-INT ist bezüglich des normalen Schreibens höher als beim PRO-ISO, da die GKS als Teilbereich der gesamten Rechtschreibleistung adressiert wird. Eine Einschränkung der Validität ist gegeben, wenn die produktive GKS-Leistung auch auf das freie Schreiben bezogen werden soll. U. a. wäre zu erwarten, dass das Generieren eigener Sätze oder Texte viele kognitive Ressourcen bindet (Sweller 1988), sodass die Aufmerksamkeit und die Verarbeitungskapazität für die Rechtschreibung reduziert wäre.

Die relativen Schwierigkeiten von Konkreta, Abstrakta und Nominalisierungen entsprechen der schon vielfach belegten Rangreihe (Bangel 2022; Betzel 2015; Rautenberg & Wahl 2024). Der Test kann diese offenbar auch erfassen (Übereinstimmungsvalidität). Überraschend ist aber, dass die NPs mit direkt vorangestelltem Artikel in diesem Test schwieriger sind als die beiden syntaktischen Kontexte mit einem Adjektivattribut. In vielen vorliegenden Studien hat sich diese Form als die leichteste herausgestellt (u. a. Rautenberg & Wahl 2024). Hier handelt es sich vermutlich um eine spezifische Besonderheit

der in diesem Text vorkommenden Phrasen. Weiterhin ist auffällig, dass die Kleinschreibung sonstiger Wörter verhältnismäßig schwierig ist.

5.5 Test zur Erfassung der Leistung bei der Rezeption der GKS beim syntaktischen Lesen (REZ-INT)

Dieser Test erfasst einen Aspekt der rezeptiven GKS-Leistung beim Lesen. Es wird getestet, ob die durch die GKS gegebene syntaktische Information registriert und korrekt für die semantische Interpretation genutzt werden kann.

Testkonstruktion. Der REZ-INT besteht aus 16 Single-Choice-Aufgaben (vgl. Abb. 5 links). In den Aufgaben werden jeweils ein Satzanfang und zwei Fortführungen des Satzes vorgegeben. Welche der beiden Fortführungen korrekt ist, entscheidet sich danach, ob das Testwort am Satzanfang groß- oder kleingeschrieben wird. Als Testwörter werden Wörter verwendet, die es – außer der GKS – in derselben Form als Nomen sowie als Verb oder Adjektiv gibt (z. B. *Mieten – mieten*, *Süßes – süßes*, *Sorgen – sorgen*). Die Idee dieses Aufgabenformats und einige der konkreten Aufgaben gehen auf Funke (2005) und Funke und Sieger (2009) zurück. Die Nutzung und Interpretation der GKS wird bei diesen Aufgaben nicht isoliert fokussiert, sondern muss integriert in den gesamten Prozess des sinnerfassenden Lesens geleistet werden.

Wie endet der Satz?

Mia jammert: „Für meine winzige Wohnung muss ich jeden Monat 1.000 Euro bezahlen. Das ist doch verrückt! Andere Mieten in dieser Stadt ...
- O ... sind gerade mal halb so hoch."
- O ... ein Haus für denselben Preis."

Elisa verspricht: „Wenn ich aus Österreich zurückkomme, bringe ich meinem Bruder etwas süßes ...
- O ... aus einer Konditorei mit."
- O ... Marzipan oder Gebäck mit."

Aisha warnt: „Immer mehr Menschen haben Angst vor Diabetes oder einem Herzinfarkt. Viele sorgen ...
- O ... sich um ihre Gesundheit, aber bewegen sich nicht genug."
- O ... um die Gesundheit könnten durch Sport überwunden werden."

Position Testwort	Anordnung Fortführungen	GKS Testwort *klein*	GKS Testwort *groß*
	Verb-Nomen		
eingebettet	*1. korrekt*	1	2
	2. korrekt	2	1
final	*1. korrekt*	1	2
	2. korrekt	2	1
	Adjektiv-Nomen		
eingebettet	*1. korrekt*	1	-
	2. korrekt	-	1
final	*1. korrekt*	-	1
	2. korrekt	1	-

Abb. 5: Beispielitems und Systematik der Testkonstruktion des REZ-INT (nach Funke & Sieger 2009)

Bei der Auswahl der Testwörter und der Konstruktion der Aufgaben wurden vier Merkmale variiert, die nach Funke und Sieger (2009) einen Einfluss auf deren Schwierigkeit haben bzw. bei denen in Analogie zu den produktiven Fehlerquellen ein Einfluss naheliegt (vgl. Abb. 5 rechts): Es gibt zwölf Aufgaben, bei denen das Testwort ein lexikalisches Verb ist, und vier Aufgaben, bei denen es ein lexikalisches Adjektiv ist.[5] Jeweils die Hälfte der Testwörter ist kleingeschrieben (syntaktisches Verb oder Adjektiv), die andere Hälfte ist großgeschrieben (syntaktisches Nomen). Kombiniert mit diesen beiden Merkmalen sind jeweils acht Testwörter in den vorgegebenen Satzanfang eingebettet (z. B. *Andere Mieten in dieser Stadt… sind gerade mal halb so hoch.*) und acht Testwörter stehen an dessen Ende,

[5] Bei Funke und Sieger (2009) werden nur syntaktische Nomen mit syntaktischen Verben kontrastiert. In dieser Studie wurden Adjektive mitaufgenommen, auch wenn sich diese im Satz anders verhalten als Verben.

also direkt vor den beiden möglichen Fortführungen (z. B. *Viele sorgen… sich um ihre Gesundheit* […]); bei acht Aufgaben ist die korrekte Satzfortführung außerdem als erste Antwortoption und bei acht Aufgaben als zweite angegeben. Die SuS benötigten zur Bearbeitung zwischen zehn und 20 Minuten.

Ergebnisse. Die acht Teilskalen und die Gesamtskala weisen eine angemessene Passung zum Testmodell auf. Die standardisierten gewichteten MNSQ-Werte liegen im Bereich T ≤ +/-2 (vgl. Tab. 4).

Auch bei diesem Test fallen die beiden Reliabilitätsmaße deutlich unterschiedlich aus (vgl. Tab. 4): Die EAP/PV-Reliabilitäten der Skalen befinden sich im ungenügenden bis unteren Bereich zwischen 0.21 und 0.45. Fasst man alle 16 Aufgaben zu einer Skala *Syntaktisches Lesen* zusammen, ergibt sich eine mittlere EAP/PV-Reliabilität von 0.515. Die Trennschärfen sind dagegen bis auf eine Teilskala sehr gut (TR ≥ .985). Die Skala mit den vier Items, bei denen zwischen einem Adjektiv und einem Nomen unterschieden werden musste, hat nur eine Trennschärfe von TR = .396. Zur Diskussion dieses Problems siehe unten.

Tab. 4: Statistische Kennwerte zu den Aspekten des REZ-INT

REZ-INT	Gewichteter Fit		Reliabilität		Schwierigkeit	
	MNSQ	*T*	*EAP/PV*	*TR*	*M*	*SD*
Testwort Verb-Nomen	[0.98 ; 1.04]	[-0.4 ; 0.9]	.451	.987	0.67	1.03
Testwort Adjektiv-Nomen	[0.98 ; 1.03]	[-0.2 ; 0.9]	.208	.396	-1.00	1.22
Testwort klein	[0.99 ; 1.05]	[-0.2 ; 1.2]	.363	.992	0.31	1.11
Testwort groß	[0.96 ; 1.03]	[-0.7 ; 0.5]	.386	.993	0.12	1,12
Testwort eingebettet	[0.99 ; 1.02]	[-0.1 ; 0.2]	.246	.990	0.28	1.00
Testwort final	[0.92 ; 1.06]	[-1.7 ; 1.4]	.432	.993	0.17	1.21
1. Fortführung korrekt	[1.00 ; 1.01]	[-0.1 ; 0.2]	.245	.989	1.01	0.98
2. Fortführung korrekt	[0.96 ; 1.04]	[-0.7 ; 0.8]	.452	.985	-0.62	1.14
Syntaktisches Lesen (gesamt)	[0,9 ; 1.00]	[-1,5 ; 1.1]	.515	.991	0.23	0.97

Die durchschnittlichen Leistungen der SuS hinsichtlich der verschiedenen Aspekte sind in Tab. 4 und Abb. 6 dargestellt. Die rezeptive Nutzung der GKS ist bei Adjektiven (M = -1.00) schwieriger als bei Verben (M = 0.67) ($F(1, 234) = 418.6$; $p < .001$; $\eta^2 = .641$). Aufgaben mit großgeschriebenen Testwörtern (M = 0.12) sind schwieriger als Aufgaben mit kleingeschriebenen Testwörtern (M = 0.31) ($F(1, 234) = 5.6$; $p = .019$; $\eta^2 = .023$). Wenn die erste angegebene Satzfortführung korrekt ist (M = 1.01), wurden die Aufgaben häufiger korrekt gelöst, als wenn die zweite Satzfortführung korrekt ist (M = -0.62) ($F(1, 234) = 458.5$; $p < .001$; $\eta^2 = .662$). Zwischen Aufgaben mit eingebettetem (M = 0.28) und finalem (M = 0.17) Testwort ergab sich kein signifikanter Unterschied ($F(1, 234) = 1.9$; $p = .172$).

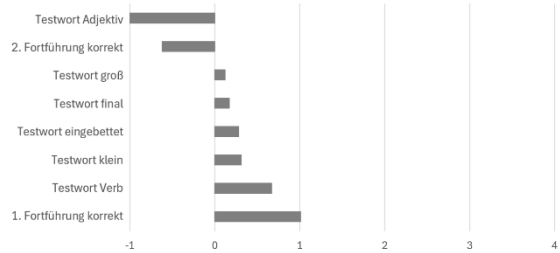


Abb. 6: Mittelwerte der Fähigkeitsparameter im REZ-INT

Bewertung. Das geschlossene Aufgabenformat des REZ-INT sichert eine hohe Durchführungsobjektivität des Messinstruments. Die EAP/PV-Reliabilitäten der einzelnen Aspekte des syntaktischen Lesens sind schlecht. Eine differenzierte Messung dieser einzelnen Aspekte wäre aber auch nur relevant, wenn damit spezifische Annahmen über die zugrundeliegenden Verarbeitungsprozesse geprüft werden sollen, die direkt mit diesen Merkmalen im Zusammenhang stehen. Betrachtet man das syntaktische Lesen insgesamt als eine Skala, ist deren Reliabilität im mittleren Bereich. Auch die Skalen dieses Tests weisen (bis auf eine Ausnahme) sehr gute Trennschärfen und eine hohe Änderungssensitivität auf, die sich in deutlichen Verbesserungen in der Evaluationsstudie (Rautenberg et al. 2025) zeigen.

Inhaltsvalidität hat der REZ-INT nur für einen rezeptiven Aspekt der GKS-Leistung, produktive Nutzungskontexte werden nicht erfasst. Aus methodischen Gründen können in den Aufgaben nur Testwörter verwendet werden, die in einer groß- und kleingeschriebenen Form existieren. Im normalen Schriftsprachgebrauch treten satzinterne Majuskeln hingegen kaum in disambiguierender Funktion auf (Mentrup 1993). Insofern ist von einer eingeschränkten Konstruktvalidität des Tests auszugehen, da er sich nur auf Fälle bezieht, in der die GKS semantische Ambiguitäten auflöst.

Wie bei den Tests, die die produktiven GKS-Kompetenzen messen, besteht auch bei REZ-INT ein deutlicher Schwierigkeitsunterschied zwischen Verben und Adjektiven. Die Proband:innen sind bei Aufgaben zu Verb-Nomen-Kontrasten häufiger dazu in der Lage, die passende Antwort auszuwählen. Dass Adjektiv-Nomen-Kontraste den SuS mehr Schwierigkeiten bereiten, kann mehrere Ursachen haben. So könnte, zumindest bei einer Testaufgabe, die kritische Einheit *Kleine* im Satzanfang *Plötzlich gibt die Kleine Emilia…* als Eigenname (*Kleine Emilia*) interpretiert worden sein. Ein weiterer Grund könnte sein, dass sich nominalisierte Adjektive syntaktisch anders verhalten als nominalisierte Verben. Einerseits können in der NP mit Adjektiven im Kern Modifikatoren beigefügt werden (z. B. *eine echt Kleine*), was ein Merkmal von Adjektiven, nicht von Nomen ist. Andererseits können sie elliptisch verwendet werden und müssen dann kleingeschrieben werden (*Im Korb liegen viele unterschiedlich große Kugeln. Ich wähle eine kleine*).

Der Effekt der Anordnung der richtigen und falschen Satzergänzung könnte auf Verschiedenes zurückzuführen sein. Es könnte ein einfacher Aufwandseffekt sein. Wenn schon die

erste Ergänzung passt, muss die zweite gar nicht gelesen werden. Dies könnte aber auch bei SuS, die die GKS nicht korrekt nutzen können, zu zufällig korrekten Lösungen führen, da sie unabhängig von der GKS ihre semantische Interpretation an die erste Ergänzung anpassen und dies subjektiv als stimmig einschätzen.

Funke und Sieger (2014), auf die das Aufgabenformat zum syntaktischen Lesen dieser Studie zurückgeht, gehen davon aus, dass es die Arbeitsgedächtniskapazität der Leser:innen übersteigen würde, wenn sie bei Aufgaben zum syntaktischen Lesen zunächst beide Antwortalternativen lesen und diese im Kopf behalten, während sie den Einleitungssatz nach der kritischen Einheit absuchen. Dieses Vorgehen ist den Autor:innen zufolge ein sehr fehleranfälliger Prozess, da auf den ersten Blick für die Leser:innen beide Antworten passend scheinen und die Auflösung der scheinbaren Mehrdeutigkeit von einem einzelnen kritischen Wort abhängt, zu dem die Leser:innen aber keinen Hinweis bekommen. Nach Funke und Sieger haben die Leser:innen nur dann eine Chance, die Aufgaben zuverlässig korrekt zu lösen, wenn sie die GKS „der kritischen Einheit bereits interpretiert haben, bevor sie die beiden Antwortalternativen lesen, dadurch auf die potentielle Mehrdeutigkeit des Satzfragments aufmerksam werden und so zum Absuchen des Satzfragments veranlasst werden“ (Funke & Sieger 2014, 4). Wenn man dies voraussetzt, dann finden die Proband:innen die korrekte Antwort „mit umso höherer Wahrscheinlichkeit […], je früher sie präsentiert wird […]“ (ebd.). Dies könnte den in der *KeGS*-Studie gefundenen Effekt der Position der richtigen Antwortalternative, zumindest für die Leser:innen, die die GKS überhaupt interpretieren, plausibel erklären.

5.6 Vergleich und Zusammenhang der Ergebnisse bei den drei Messverfahren

Vergleicht man die Gesamtleistungen bei den drei Messverfahren, ergibt sich folgende Rangfolge: Beim PRO-ISO erreichten die SuS die höchsten Werte (M = 1.67, s = 0.99), gefolgt vom PRO-INT (M = 1.16, s = 0.67), beim REZ-INT schnitten sie am schlechtesten ab (M = 0.23, s = 0.97). Die Schwierigkeit der drei Tests unterscheidet sich deutlich (F(1.7, 397.1) = 285.4; p < .001; η^2 = .549). Bei den beiden produktiven Tests zeigen sich dieselben Schwierigkeitsunterschiede für Konkreta (PRO-ISO: M = 2.87, s = 1.47; PRO-INT: M = 2.14, s = 0.84; F(1, 234) = 80.0, p < .001, η^2 = .255) und Abstrakta (PRO-ISO: M = 1.90, s = 1.72; PRO-INT: M = 1.68, s = 1.20; F(1, 234) = 5.1, p = .025, η^2 = .021). Bei den Nominalisierungen ist kein Unterschied festzustellen (PRO-ISO: M = -0.64, s = 1.29; PRO-INT: M = -0.48, s = 1.28; F(1, 234) = 3.3, p = .072).

Während beim PRO-ISO die Aufmerksamkeit ausschließlich auf die GKS-Entscheidung gerichtet werden kann, beanspruchen beim PRO-INT viele weitere Anforderungen die Verarbeitungsressourcen der SuS, sodass sie beim PRO-ISO erwartungsgemäß besser abschneiden als beim PRO-INT. Die Schwierigkeiten beim REZ-INT lassen sich möglicherweise darauf zurückführen, dass die GKS normalerweise gar nicht oder nur selten zur Disambiguierung herangezogen werden muss. Zudem kann das schlechte Abschneiden der SuS mit der zusätzlichen Herausforderung zusammenhängen, sich nach dem Lesen und Verstehen der ersten Fortführung eines Satzes – völlig unabhängig von der Rechtschreibung – überhaupt auf eine zweite neue Satzstruktur und -bedeutung einzulassen. Selbst bei geübten

Leser:innen, die die GKS sicher für sich zu nutzen wissen, ist dies kognitiv herausfordernd (garden path effect).

Trotz der deutlichen Schwierigkeitsunterschiede zwischen den drei Testverfahren gibt es signifikante Korrelationen zwischen den Messwerten. Die Gesamtwerte der produktiven Tests PRO-ISO und PRO-INT hängen deutlich zusammen ($r = .673$, $p < .001$). Auch zwischen den Messwerten zur Großschreibung der Konkreta, Abstrakta und Nominalisierungen zeigen sich signifikante Korrelationen (vgl. Tab. 5). Zwischen dem rezeptiven REZ-INT und den anderen beiden Tests bestehen schwache Korrelationen (REZ-INT und PRO-ISO: $r = .393$, $p < .001$; REZ-INT und PRO-INT: $r = .359$, $p < .001$).

Tab. 5: Korrelationen zwischen den Leistungen in den drei Tests

Korrelationen		PRO-INT				REZ-INT
		Konkreta	*Abstrakta*	*Nominal.*	*Gesamt*	*Gesamt*
PRO-ISO	*Konkreta*	.539**	.472**	.330**	.599**	.272**
	Abstrakta	.455**	.527**	.414**	.639**	.341**
	Nominalisierungen	.120*	.265**	.410**	.662**	.340**
	Gesamt	.476**	.519**	.439**	.673**	.393**
REZ-INT	*Gesamt*	.153*	.203**	.333**	.359**	

Alle drei Tests beziehen sich zwar auf die Messung der GKS-Leistung, messen jedoch unterschiedliche Aspekte, was sich in den moderaten Korrelationen widerspiegelt. Während PRO-ISO und PRO-INT produktive Kompetenzen in den Blick nehmen, werden über REZ-INT rezeptive Kompetenzen beim syntaktischen Lesen erfasst. PRO-ISO und PRO-INT wiederum unterscheiden sich darin, ob die GKS im Fokus der Aufmerksamkeit steht (PRO-ISO) oder ob sie in den Schreibprozess integriert erfolgt und das Arbeitsgedächtnis hier durch andere Prozesse (u. a. Beachtung der Orthographie auf Wortebene, (fein)motorische Prozesse bei der Verschriftung, Verarbeitung des Wortschatzes) zusätzlich belastet ist (PRO-INT). Es war entsprechend zu erwarten, dass die GKS den SuS im Lückendiktat (PRO-INT) weniger zuverlässig gelingt als in PRO-ISO.

Der statistische Zusammenhang zwischen den Leistungen in den verschiedenen Nutzungskontexten erlaubt aber keine Schlussfolgerungen über die zugrundeliegenden Prozesse der Informationsverarbeitung, die GKS-Leistungen ermöglichen. Es bleibt offen, ob die Kompetenz, in vorgegebenen Sätzen korrekte GKS-Entscheidungen zu treffen, die GKS beim Schreiben positiv beeinflusst oder umgekehrt. Plausibel ist es, anzunehmen, dass die in PRO-ISO erforderliche Tätigkeit näher an der Tätigkeit der Überarbeitung von Texten im Schreibprozess liegt (in der Überarbeitungsphase wird der geschriebene Text u. a. auf orthographische Korrektheit überprüft) und dass diese Fähigkeit auch die Überarbeitungskompetenzen bezüglich GKS beeinflussen könnten. Die mit PRO-INT verbundene Tätigkeit hingegen erfolgt in den Schreibprozess integriert (vgl. Bredel & Hlebec 2015).

6 Diskussion

Mit REZ-INT wurde ein Messverfahren für die rezeptive GKS-Kompetenz vorgestellt, das auf Funke (2005) und Funke und Sieger (2014) zurückgeht. In bisherigen Interventionsstudien wurden (mit Ausnahme von Funke et al. 2013) ausschließlich produktive GKS-Leistungen erhoben und damit nur Teilaspekte von Großschreibkompetenz. Insofern decken die drei vorgestellten Testverfahren drei zentrale Nutzungskontexte der GKS ab. Durch die vollständige und systematische Kombination der möglichen Fälle der GKS in den beiden produktiven Tests wird zusätzlich auch der inhaltliche Anforderungsbereich der GKS erfasst. So kann man sich in Evaluationsstudien über eine große Bandbreite an Aspekten von GKS-Leistungen ein differenzierteres Bild machen und Hinweise auf vorhandene und noch auszubauende Kompetenzen geben.[6]

Bei Evaluationsstudien, die sich auf didaktische Ansätze beziehen und damit schulpraktische Implikationen haben, ist es notwendig, auch die Wirkung auf die Leistungen in einzelnen Teilbereichen (z. B. die Großschreibung von Nominalisierungen) zu messen. Da jedes Testitem aber immer gleichzeitig viele Merkmale hat, die die Leistung beeinflussen können (z. B. Wortart, syntaktischer Kontext, morphologische Komplexität), ist die Wirkung eines einzelnen Merkmals grundsätzlich mit den anderen Merkmalen konfundiert. Deshalb erfolgte die Itemkonstruktion systematisch entlang eines Schemas, bei dem alle möglichen Kombinationen der relevanten Merkmale verwendet wurden. Dadurch decken die Items zu *einem* Merkmal alle (möglichen) Kombinationen zu den jeweils anderen Merkmalen ab, sodass die Wirkung der anderen Merkmale als kontrolliert betrachtet werden kann. Es gibt bspw. Items zu Abstrakta und Nominalisierungen, die jeweils in allen syntaktischen Kontexten (+/- Artikel, +/- Adjektiv) präsentiert wurden. Wenn sich nun Leistungsunterschiede zwischen Abstrakta und Nominalisierungen zeigen, kann dies nicht an den syntaktischen Kontexten liegen, weil die Konfundierung durch die systematischen Kombinationen kontrolliert wurde.

Ein Nebeneffekt dieses Vorgehens ist, dass die Schwierigkeiten der Items zu einem Merkmal (einer Skala) relativ stark streuen. Die Schwierigkeiten der nominalisierten Items ergeben sich z. B. auch aus dem syntaktischen Kontext, also ob die Nominalisierung in einer NP mit oder ohne Artikel und mit oder ohne Adjektivattribut präsentiert wird. Dies führt bei der Zusammenfassung der Items auf einer Skala zwangsläufig dazu, dass die Items nicht vollständig homogen sind, was sich bei den Skalenanalysen in geringeren Reliabilitäten zeigt. Eine Erhöhung der Homogenität (und Reliablität) wäre durch eine Konstanthaltung der jeweils anderen Merkmale möglich (z. B. alle Items mit Nominalsierungen werden im selben syntaktischen Kontext, bspw. mit Artikel und ohne Adjektivattribut, präsentiert). Dies hätte dann aber wiederum eine Einschränkung der Validität zur Folge, weil nur ein Teil der vorkommenden Fälle abgedeckt wären. Den bei der Testkonstruktion häufig auftretenden Trade-off zwischen Validität und Reliabilität haben Holland und Gottfredson (1976) als Bandbreiten-Fidelitäts-Dilemma bezeichnet.

[6] Nicht erhoben wurde die Fähigkeit, GKS-Entscheidungen nachträglich zu begründen (vgl. Wahl & Rautenberg 2020; Rautenberg 2021).

Im vorliegenden Fall hat die Entscheidung für eine große Bandbreite bei den Tests zu nur mittleren bis teilweise auch schlechten EAV/PV-Reliabilitäten geführt. Dies wäre insbesondere dann kritisch, wenn die Messwerte einzelner SuS interpretiert und pädagogische Konsequenzen haben würden. Da die Testverfahren in einer Evaluationsstudie eingesetzt und nur die Mittelwerte von Teilstichproben der SuS verglichen und interpretiert wurden, sind diese Reliabilitätsmaße nicht gleichermaßen relevant. Bei dieser Verwendung sind vor allem die sehr guten Trennschärfen wichtig, da sie eine gute Differenzierung in unterschiedliche Fähigkeitsniveaus anzeigen.

In den Tests wurden aus methodischen Gründen keine Sätze oder Texte verwendet, die die SuS selbst produziert haben. Es wurden nur Fremdtexte verwendet. Das „integrierte Prozesswissen" (Bredel 2013, 110) wurde im Rahmen eines Lückendiktats ebenfalls an vorgegebenen Sätzen ermittelt. Es kann jedoch davon ausgegangen werden, dass die Bekanntheit des Textes die Aufgabenanforderung beeinflusst, etwa weil eigene Texte keinen unbekannten Wortschatz aufweisen und das Textverständnis vorausgesetzt werden kann. Dies kann eine Orientierung im Text erleichtern.

Allerdings entlasten vorgegebene, fremde Schreibungen das Arbeitsgedächtnis, da der Schreibprozess sich auf das Verschriften oder Auswerten vorgegebener Sätze/Texte beschränkt. Die Distanzierung, die für eine sprachliche Analyse notwendig ist, könnte bei fremden Texten leichter fallen als bei selbst verfassten.

Zudem sind beim Verfassen eigener Texte die kognitiven Ressourcen im Schreibprozess durch simultan zu bewältigende Aufgaben stark belastet, was u. a. dazu führen kann, dass die Zahl orthographischer Fehler in selbst verfassten Texten insgesamt höher ist als in Anforderungssituationen, in denen der Fokus auf der orthographischen Korrektheit liegt; zumindest ist dies anzunehmen, wenn es sich um den ersten Textentwurf handelt. Bei Textüberarbeitungen hingegen ist davon auszugehen, dass die GKS nicht prozessintegrierend vorgenommen wird, sondern – ähnlich wie in PRO-ISO – nachträglich und isoliert.

Methodisch besteht bei der Erhebung von GKS-Leistungen in dieser Hinsicht ein Validitäts-Reliabilitäts-Dilemma. Freie Texte als Messinstrumente weisen eine höhere Validität hinsichtlich des Nutzungskontextes auf, sind aber aufgrund der nicht kontrollierbaren Frequenz des Auftretens der unterschiedlichen Fälle der GKS sehr unreliabel und subjektiv und daher auch weniger valide auf der Dimension der inhaltlichen Abdeckung des Konstrukts. Dies führt zu einer geringen Vergleich- und Messbarkeit der Leistungen, die jedoch für Vergleichsstudien sichergestellt sein muss. Für Vergleichsstudien scheinen freie Schreibungen daher ungeeignet. Misst man GKS-Kompetenzen allerdings nur über die Korrektur oder Verschriftung von Fremdtexten und damit in einer künstlichen Testsituation, wird eine allgemeine GKS-Kompetenz ggfs. nicht präzise genug erfasst, denn das Verfassen freier Texte stellt schließlich die ‚normale' Schreibsituation dar. Das Dilemma kann nicht vollständig aufgelöst werden. Wir sprechen uns dafür aus, in empirischen Studien, die die GKS-Kompetenzen von Lerner:innen erfassen, auf vergleichbare Testverfahren zurückzugreifen, um die genannten methodischen Schwierigkeiten zu vermeiden.

Anders verhält es sich jedoch, wenn es um eine Textanalyse im Rahmen einer Einzelfalldiagnostik oder Lernstandserhebung geht. Hier kann und sollte zusätzlich eine kriteriengeleitete Analyse von freien Schreibungen durchführt werden, die Aufschluss darüber geben kann, welche Schwerpunkte bei einer Förderung der GKS-Kompetenzen sinnvollerweise gesetzt werden sollten.

Literatur

Amtliches Regelwerk (2024). Deutsche Rechtschreibung: Regeln und Wörterverzeichnis. Aktualisierte Fassung des amtlichen Regelwerks entsprechend den Empfehlungen des Rats für deutsche Rechtschreibung 2024. IDS Mannheim. https://grammis.ids-mannheim.de/rechtschreibung/

Bangel, M. (2022). Potentiale einer syntaxbasierten Vermittlung der satzinternen Großschreibung in Jahrgang 5 und mögliche vermittlungsunabhängige Einflussfaktoren. In H. Hlebec & S. Sahel (Hrsg.), *Orthographieerwerb im Übergang* (119–145). Erich Schmidt. https://doi.org/10.37307/b.978-3-503-20650-6.06

Betzel, D. (2015). Zum weiterführenden Erwerb der satzinternen Großschreibung: Eine leistungsgruppendifferenzierte Längsschnittstudie in der Sekundarstufe I. Schneider.

Bredel, U. (2006). Die Herausbildung des syntaktischen Prinzips in der Historiogenese und in der Ontogenese der Schrift. In U. Bredel & H. Günther (Hrsg.), *Orthographietheorie und Rechtschreibunterricht* (139–164). Niemeyer.Bredel, U. (2013). Sprachbetrachtung und Grammatikunterricht. 2. Aufl. Schöningh.

Bredel, U. (2013). Sprachbetrachtung und Grammatikunterricht. 2. Aufl. UTB.

Bredel, U. & Hlebec, H. (2015). Kommasetzung im Prozess. *Praxis Deutsch* 254, 36–43.

Brucher, L., Ugen, S. & Weth, C. (2020). The impact of syntactic and lexical trainings on capitalization of nouns in German in grade five. *L1 Educational Studies in Language and Literature* 20, 1–23. https://doi.org/10.17239/L1ESLL-2020.20.01.01

Eisenberg, P. (2020). Der Satz. Grundriss der deutschen Grammatik (5. Aufl.). Metzler. https://doi.org/10.1007/978-3-476-05096-0

Fuhrhop, N. & Romstadt, J. (2021). Orthographiefehler im Abitur: Eine sprachwissenschaftliche Bestandsaufnahme. In M. Kepser, S. Schallenberger & H.-G. Müller (Hrsg.), *Neue Wege des Orthografieerwerbs: Forschung – Vermittlung – Reflexion* (189–208). Lemberger.

Funke, R. (1995). Das Heben des Wortschatzes: Nomen im Kontext sehen. *Praxis Deutsch* 22(129), 57–60.

Funke, R. (2005). Sprachliches im Blickfeld des Wissens. Niemeyer.

Funke, R. & Sieger, J. (2009). Die Nutzung von orthographischen Hinweisen auf syntaktische Strukturen und ihre Bedeutung für das Leseverständnis. Empirische Daten und

didaktische Folgerungen. *Didaktik Deutsch: Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur* 14, 31–53.

Funke, R. & Sieger, J. (2014). Grammatische Strukturen als Lerngegenstand im Deutschunterricht: Um welche Art von Lernen geht es? In U. Bredel & C. Schmellentin (Hrsg.), *Welche Grammatik braucht der Grammatikunterricht?* (1–22). Schneider.

Funke, R., Wieland, R., Schönenberg, S. & Melzer, F. (2013). Exploring syntactic structures in first-language education: Effects on literacy-related achievements. *L1 Educational Studies in Language and Literature* 13, 1–24.

Gaebert, D.-K. (2012). Zur Didaktik der satzinternen Großschreibung im Deutschen. Wortartbezogene Umwege und syntaktische Katalysatoren. Peter Lang.

Gallmann, P. (1997). Konzepte der Nominalität. In G. Augst, K. Blüml, D. Nerius & H. Sitta (Hrsg.), *Zur Neuregelung der deutschen Orthographie* (209-242). De Gruyter.

Günther, H. & Nünke, E. (2005). Warum das Kleine großgeschrieben wird, wie man das lernt und wie man das lehrt. Kölner Beiträge zur Sprachdidaktik, 1.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage Publications, Inc.

Holland, J. L., & Gottfredson, G. D. (1976). Using a typology of persons and environments to explain careers: Some extensions and clarifications. *Journal of Counseling Psychology, 23*, 423–432.

Hückmann, A., Siegel, V., Rautenberg, I. & Wahl, S. (eingereicht). Acquiring German noun capitalization with and without explicit syntactic knowledge – two digital tutorials for secondary school.

Hückmann, A. (in Vorbereitung). Lernprozesse beim Erwerb der satzinternen Großschreibung (Arbeitstitel).

Maas, U. (1992). Grundzüge der deutschen Orthographie. De Gruyter. https://doi.org/10.1515/9783111376974

Mangelschots, K., Ugen, S. & Weth, C. (2023). Profiles of poor and good spellers in German noun capitalization. *L1-Educational Studies in Language and Literature* 23(1), 1–21.

Mentrup, W. (1993). Wo liegt eigentlich der Fehler? Zur Rechtschreibreform und ihren Hintergründen. Klett.

Müller, H.-G. (2016). Der Majuskelgebrauch im Deutschen. Groß- und Kleinschreibung theoretisch, empirisch, ontogenetisch. De Gruyter.

Pießnack, C. & Schübel, A. (2005). Untersuchungen zur orthographischen Kompetenz von Abiturientinnen und Abiturienten im Land Brandenburg. In H. Giest (Hrsg.), *LLF- Berichte* 20 (50–72). Universitätsverlag.

Rautenberg, I. (2021). Die Berücksichtigung syntaxbasierter Ansätze zur Großschreibung in Lehrwerken für die Sekundarstufe. Eine exemplarische Analyse. *Der Deutschunterricht* 3, 49–60.

Rautenberg, I. & Wahl, S. (2024). Welche sprachstrukturellen Faktoren beeinflussen die Großschreibung von Schüler*innen im Deutschen? *Zeitschrift Für Sprachlich-Literarisches Lernen Und Deutschdidaktik* 4, 1–27. https://doi.org/10.46586/SLLD.Z.2024.11376

Rautenberg, I., Wahl, S., Siegel, V. & Hückmann, A. (2025). Syntaxbasierte Vermittlung der satzinternen Großschreibung mit Online-Kursen – Evaluation impliziter und expliziter didaktischer Ansätze in der Sekundarstufe. *Didaktik Deutsch* 59, 33–56.

Röber-Siekmeyer, C. (1999). Ein anderer Weg zur Groß- und Kleinschreibung. Ernst Klett.

Steinig, W. & Betzel, D. (2014). Schreiben Grundschüler heute wirklich schlechter als vor 40 Jahren? Texte von Viertklässlern aus den Jahren 1972, 2002 und 2012. In A. Plewnia & A. Witt (Hrsg.), *Sprachverfall. Dynamik – Wandel – Variation* (353–371). De Gruyter. https://doi.org/10.1515/9783110343007.353

Sweller, J. (1988): Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science* 12, 257–285.

Wahl, S. & Rautenberg, I. (2020). Explizites orthographisches Wissen von Grundschulkindern über die satzinterne Großschreibung. In I. Rautenberg (Hrsg.), *Evidenzbasierte Forschung zum Schriftspracherwerb* (128–145). Schneider.

Wahl, S., Rautenberg, I. & Helms, S. (2017). Evaluation einer Didaktik zur satzinternen Großschreibung. *Didaktik Deutsch* 42, 32–52.

Weth, C., Dording, C., Klasen, L., Michel, F., Funke, R. & Ugen, S. (2024). Effects of parallel syntactic training in French plural spelling and German noun capitalization. *Morphology* 34, 1–29. https://doi.org/10.1007/s11525-024-09420-3

7 Korrespondenzangaben

Dr. Stefan Wahl, Pädagogische Hochschule Freiburg, Institut für Psychologie

Prof. Dr. Iris Rautenberg, Pädagogische Hochschule Ludwigsburg, Institut für deutsche Sprache und Literatur

Alicia Hückmann, Pädagogische Hochschule Ludwigsburg, Institut für deutsche Sprache und Literatur

Dr. Vanessa Siegel, Pädagogische Hochschule Freiburg, Institut für Psychologie

Korrespondenz an: wahl@ph-freiburg.de