Z Erziehungswiss (2019) 22:1281–1312
https://doi.org/10.1007/s11618-019-00910-3

ZfE

ALLGEMEINER TEIL

Schreibkompetenzen im Fach Englisch in der gymnasialen Oberstufe

Olaf Köller · Johanna Fleckenstein · Jennifer Meyer · Anna Lara Paeske · Maleika Krüger · Andre A. Rupp · Stefan Keller

Online publiziert: 19. November 2019
© Der/die Autor(en) 2019

Zusammenfassung Produktive Sprachkompetenzen im Fach Englisch sind bislang in Deutschland nur wenig untersucht worden. Daher wurden in der vorliegenden Untersuchung mit zwei Messzeitpunkten Kompetenzen im argumentativen und sachorientierten Schreiben von $N = 838$ Schülerinnen und Schülern in der 11. Jahrgangsstufe

Die Daten der vorliegenden Studie stammen aus dem Projekt „Measuring English Writing at Secondary Level (MEWS)“, das von der Deutschen Forschungsgemeinschaft (DFG; GZ: KO1513/12-1) und vom Schweizerischen National Fonds (SNF: Fördernr. 100019L162675) gefördert wird.

Prof. Dr. O. Köller (✉) · Dr. J. Fleckenstein · J. Meyer · A. L. Paeske
Abteilung Erziehungswissenschaft und Pädagogische Psychologie, IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Olshausenstraße 62, 24098 Kiel, Deutschland
E-Mail: koeller@ipn.uni-kiel.de

Dr. J. Fleckenstein
E-Mail: fleckenstein@ipn.uni-kiel.de

J. Meyer
E-Mail: jmeyer@ipn.uni-kiel.de

A. L. Paeske
E-Mail: paeske@ipn.uni-kiel.de

M. Krüger · Prof. Dr. S. Keller
Pädagogische Hochschule, Englischdidaktik und ihre Disziplinen, FHNW Campus, Fachhochschule Nordwestschweiz FHNW, Hofackerstr. 30, 4132 Muttenz, Schweiz

M. Krüger
E-Mail: maleika.krueger@fhnw.ch

Prof. Dr. S. Keller
E-Mail: stefan.keller@fhnw.ch

Dr. A. A. Rupp
Educational Testing Service (ETS), 660 Rosedale Road, Thurstone Building, 03-T, Princeton, NJ 08541, USA
E-Mail: arupp@ets.org

des achtjährigen Gymnasiums untersucht. Zusätzlich wurden rezeptive Kompetenzen (Hören und Lesen) berücksichtigt. Mit Blick auf die Ziele der gymnasialen Oberstufe (Erreichen des Niveaus B2 des Gemeinsamen Europäischen Referenzrahmens für Sprachen; GER) konnte gezeigt werden, dass bereits ein Jahr vor Erreichen des Abiturs rund 60 % der getesteten Schülerinnen und Schüler das Niveau B2 oder höher im Schreiben erreichten. Weiterhin belegen die Analysen signifikante Kompetenzzuwächse im Laufe eines Schuljahres und zeigen deutliche Unterschiede zwischen den Profilen (sprachlich vs. naturwissenschaftlich vs. gesellschaftswissenschaftlich vs. sonstige), die die Schülerinnen und Schüler in der gymnasialen Oberstufe belegt haben; hier zeigen Jugendliche in den Sprachprofilen signifikant höhere Leistungen. Die Ergebnisse werden mit Bezug auf normative Zielvorgaben der gymnasialen Oberstufe diskutiert.

Schlüsselwörter Fremdsprachenkompetenzen · Gymnasiale Oberstufe · Gemeinsamer Europäischer Referenzrahmen für Sprachen · Ziele schulischen Lernens

Writing skills in English as a foreign language in upper secondary school

Abstract There is a substantial lack of studies in productive skills of German students in English as a foreign language. Based on this research gap, the present repeated measurement study evaluated the competence to write argumentative essays and synthesis texts of $n = 838$ 11th graders from upper secondary schools while also collecting data on receptive skills (listening and reading comprehension). Findings show that a substantial number of students (approximately 60%) reach level B2 or higher of the Common European Framework of Reference for Languages one year before leaving upper secondary school, which is the official standard for upper secondary school writing. Furthermore, analyses show that all foreign language skills increased over the course of one school year and that students with different educational emphases in their school tracks (i.e., with a primary focus on languages vs. natural science vs. social science vs. other subjects) differ significantly in their writing skills. Specifically, students who pursue language-centered studies clearly outperformed all other students as expected. Findings are discussed with respect to normative performance expectations of upper secondary schooling.

Keywords Common European Framework of Reference for Languages · Foreign language skills · Normative goals of schooling · Upper secondary school

1 Einleitung

Hinreichende Kompetenzen im Bereich des argumentativen und materialbasierten Schreibens gelten ebenso wie rezeptive Kompetenzen als ein zentrales Ziel eines modernen, kommunikativ ausgerichteten Englischunterrichts in der gymnasialen Oberstufe. Dazu hat die Kultusministerkonferenz (KMK 2012) entsprechende Bildungs-

standards für die fortgeführte Fremdsprache (Englisch/Französisch) verabschiedet, in denen sie fünf kommunikative Kompetenzen (Hör-/Hörsehverstehen, Leseverstehen, Schreiben, Sprechen und Sprachmittlung) beschreibt und Zielerwartungen in Bezug auf den Gemeinsamen Europäischen Referenzrahmen für Sprachen (GER, Europarat 2001) formuliert. Als Zielgröße für sprachproduktive Kompetenzen am Ende der gymnasialen Oberstufe, die im Vordergrund des vorliegenden Aufsatzes stehen, wird das GER-Niveau B2 definiert. Für rezeptive Kompetenzen werden teilweise Leistungen auf dem Niveau C1 erwartet.

Anders als im Falle der Sekundarstufe I, in der zumindest die rezeptiven Fremdsprachenkompetenzen in den Fächern Englisch und Französisch am Ende der 9. Jahrgangsstufe überprüft werden (zuletzt in 2015, vgl. Stanat et al. 2016), gibt es in der KMK keinerlei Bestreben, standardbasierte Überprüfungen der Englisch- und Französischleistungen in der gymnasialen Oberstufe zu initiieren. Dies überrascht vor allem im Falle von Englisch als *lingua franca*, die sich als Standardsprache in vielen akademischen Berufen und Wissenschaftsdisziplinen etabliert hat (Keller 2013). Trotz dieser fehlenden Initiativen der Länder bzw. der KMK existieren seit einigen Jahren große empirische Studien, die sich zumindest mit rezeptiven Leistungen in der gymnasialen Oberstufe auseinandersetzen (zuletzt Leucht et al. 2016). Es fehlen aber weitgehend Arbeiten zu sprachproduktiven Leistungen. Dies ist nicht zuletzt dem Umstand geschuldet, dass schriftliche und mündliche Leistungen einen höheren Aufwand bei der Testung, vor allem bei der Auswertung, erfordern. Allerdings wurden inzwischen im internationalen Kontext Verfahren entwickelt, um sprachproduktive Kompetenzen ökonomisch zu erheben und mit dem Durchbruch im Bereich der maschinellen Auswertung von Texten bestehen Möglichkeiten, sprachproduktive Leistungen von großen Schülerzahlen zeitnah, reliabel und valide zu bewerten (vgl. hierzu Rupp et al. 2019). In der vorliegenden Arbeit nutzen wir die großen technischen Fortschritte bei der Testung und Auswertung von Schreibkompetenzen und berichten die Ergebnisse aus einer Studie, in der über 800 Gymnasiastinnen und Gymnasiasten der 11. Jahrgangsstufe im achtjährigen Gymnasium Schleswig-Holsteins zu zwei Erhebungszeitpunkten insgesamt vier Schreibaufgaben sowie Aufgaben zum Hör- und Leseverstehen bearbeiteten. Die geschriebenen Texte wurden sowohl von geschulten menschlichen Beurteilerinnen und Beurteilern als auch computerbasiert (maschinell) ausgewertet. Ein Standard-Setting diente dazu, die Qualität der produzierten Texte und somit die Schreibkompetenzen der Schülerinnen und Schüler auf den GER-Stufen abzubilden. Als Folge lässt sich erstmals GER-basiert die Frage beantworten, wie gut – gemessen an den normativen Zielsetzungen der gymnasialen Oberstufe – die Schreibkompetenzen ausfallen. Zusätzlich wird der Frage nachgegangen, wie hoch die Qualität der Schreibleistungen von Schülerinnen und Schülern in Abhängigkeit von unterschiedlichen individuellen Schwerpunktsetzungen (Profilen; s. unten) in der Oberstufe ausfällt und wie hoch die Zuwächse der Schreibkompetenzen im Laufe eines Schuljahres sind. Ergänzend zu diesen Kernfragestellungen sollen Ergebnisse präsentiert werden, welche die hohe psychometrische Qualität des hier verwendeten Paradigmas zur Erfassung von Schreibkompetenzen belegen.

2 Theoretischer Hintergrund

Im Folgenden sollen zunächst existierende Befunde zu den Fremdsprachenkompetenzen von Schülerinnen und Schülern in der gymnasialen Oberstufe berichtet werden. Daran schließen sich Ausführungen über Herausforderungen bei der Feststellung von produktiven fremdsprachlichen Kompetenzen an. Dazu wird auch die Konzeptualisierung von fremdsprachlichen Schreibkompetenzen, die in den Bildungsstandards der gymnasialen Oberstufe ausgedrückt ist, dargestellt und bezogen auf ihre Operationalisierbarkeit analysiert. Es folgen Ausführungen zur Profiloberstufe, die in mehreren Bundesländern das System der Grund- und Leistungskurse abgelöst hat. Zum Abschluss des Theorieteils werden die Fragestellungen der vorliegenden Untersuchung präsentiert.

2.1 Fremdsprachenkompetenzen in der Sekundarstufe II im GER

Die 2012 von der KMK verabschiedeten Bildungsstandards für die fortgeführte Fremdsprache für die Abiturprüfung lehnen sich in ihren Zielerwartungen eng an den GER an und definieren die Niveaus B2 und C1 als normative Vorgaben für Schülerinnen und Schüler, die Englisch oder Französisch in der gymnasialen Oberstufe fortführen. Für Französisch fehlen unseres Wissens groß angelegte Schulleistungsstudien in der Sekundarstufe II komplett, ein besseres Bild ergibt sich im Fach Englisch. Hier existieren Arbeiten, in denen rezeptive Leistungen (Lese- und Hörverstehen, teilweise ergänzt um Grammatikkenntnisse) in der gymnasialen Oberstufe auf dem GER abgebildet wurden (vgl. Köller et al. 2004; Jonkmann et al. 2007; Jonkmann et al. 2010; sowie Leucht et al. 2016). Die älteren Arbeiten beschränken sich dabei auf Ergebnisse, die mit einer in den 1980er Jahren entwickelten Version des *Tests of English as a Foreign Language* (TOEFL) erzielt wurden. Für diesen Test haben Tannenbaum und Wylie (2005) mit Hilfe eines Standard-Setting die Verlinkung zum GER vorgenommen. Die neuere Arbeit von Leucht et al. (2016) nutzt dagegen Tests zum Hör- und Leseverstehen, die in Deutschland zur Überprüfung der Erreichung der länderübergreifenden Bildungsstandards im Fach Englisch am Ende der Sekundarstufe I entwickelt wurden (vgl. Köller et al. 2010) und im deutschen Kontext unter Einbindung europäischer Expertinnen und Experten mit dem GER verlinkt wurden (Harsch et al. 2010). Die Items haben sich auch für den Einsatz in der Sekundarstufe II bewährt (Leucht et al. 2015). Die älteren Arbeiten ebenso wie die Studie von Leucht et al. (2016) zeichnen ein relativ optimistisches Bild der rezeptiven Englischleistungen am Ende der gymnasialen Oberstufe. In der Untersuchung von Leucht et al. (2016) wurden 3775 Schülerinnen und Schüler in Schleswig-Holstein am Ende der gymnasialen Oberstufe (13. Jahrgangsstufe im neunjährigen Gymnasium) getestet. Insgesamt 38 % der Schülerinnen und Schüler besuchten ein allgemeinbildendes, die übrigen ein berufliches Gymnasium. Im allgemeinbildenden Gymnasium, in dem Englisch vierstündig auf erhöhtem Anforderungsniveau unterrichtet wird, erreichten im Lesen rund 80 % der Schülerinnen und Schüler die Niveaus B2 (57,8 %) und C1 (22,4 %). Im Hörverstehen lagen die Zahlen bei 59,9 (B2) und 36,6 % (C1). Leistungskursschülerinnen und -schüler an beruflichen Gymnasien (fünfstündig auf erhöhtem Anforderungsniveau) erreichten

folgende Leistungen: Lesen: B2/58,6 %; C1/11,5 %; Hören: B2/72,0 %; C1/18,6 %. Deutlich schwächer fielen die Leistungen in den Grundkursen aus. Insgesamt kann aber festgehalten werden, dass die normativen Erwartungen des Englisch-Unterrichts (B2/C1) von erheblichen Anteilen der Schülerinnen und Schüler am Ende der Sekundarstufe II im rezeptiven Bereich erreicht werden.

2.2 Konzeptualisierung von fremdsprachlichen Schreibkompetenzen in der gymnasialen Oberstufe

Während in der Sekundarstufe I vor allem das Schreiben zu persönlichen Themen sowie von einfachen Berichten im Mittelpunkt steht (KMK 2004), erweitert sich in der gymnasialen Oberstufe das Spektrum der angezielten Textsorten in Richtung auf das argumentative und das materialbasierte Schreiben (KMK 2012). In den Bildungsstandards der KMK für die fortgeführte Fremdsprache (Abitur) werden zwei Funktionen des fremdsprachigen Schreibens besonders betont. Erstens die persuasive oder rhetorische Funktion des Schreibens, wobei es darum geht, ein Thema zu analysieren, sich eine Meinung dazu zu bilden und eigene Argumente in der Fremdsprache so zu präsentieren, dass sie für Leserinnen und Leser überzeugend wirken. Die Jugendlichen sollen schon auf grundlegendem Anforderungsniveau fähig sein, sich argumentativ mit unterschiedlichen Positionen auseinanderzusetzen, sowie Informationen in ihren Texten strukturiert und kohärent vermitteln zu können (KMK 2012). Auf dem erhöhten Anforderungsniveau werden die Dimensionen der Adressatenorientierung sowie Textsortenorientierung besonders betont, wobei die Lernenden in den eigenen Textproduktionen situationsangemessen und adressatengerecht handeln und dabei die Konventionen der jeweiligen Textsorte beachten sollen (KMK 2012). Durch diese Kompetenzbeschreibungen ist das Genre des *Argumentive Essay* impliziert, auch wenn es in den Standards der KMK nicht explizit genannt wird.

Die zweite Funktion des Schreibens, die in den Bildungsstandards der Oberstufe zentral gesetzt ist, könnte man als „materialbasiert" oder auch „integrativ" beschreiben. Das kann z. B. bedeuten, eine Textsorte in eine andere umzuwandeln, z. B. indem „diskontinuierliche Vorlagen in kontinuierliche Texte" umgeschrieben oder auch Texte „für eine andere Zielgruppe" adaptiert werden (KMK 2012). Dabei sind komplexe Sprachproduktionsmodelle impliziert, wobei unterschiedliche Sprachkompetenzen und Verarbeitungsmodi orchestriert und integriert werden müssen, etwa, wenn Texte zuerst gelesen oder gehört werden müssen, um darauf aufbauend einen eigenen Text zu verfassen. Die Produkte solch integrativer Sprachleistungen werden auch *Synthesis Texts* genannt, wobei die Sprach*produktion* auf vorhergehender Sprach*rezeption* beruht und die beiden verbunden werden müssen (van Ockenburg et al. 2016). Dies ist kongruent mit modernen psychologischen Modellen der Schreibkompetenz, in denen das Lesen als integrativer Teil des Schreibens betrachtet wird (vgl. dazu im Detail Schoonen 2019).

Durch diese Konzeptualisierung in den Bildungsstandards wird auch klar, dass Schreiben in der Oberstufe im Sinne von *Composition* verstanden werden muss, wobei Oberflächen- und Tiefenmerkmale von Texten integriert werden müssen: „Composing involves the combining of structural sentence units into a more-or-

less unique, cohesive and coherent larger structure (as opposed to lists, forms, etc.). A piece of writing which implicates composing contains surface features which connect the discourse and an underlying logic of organization which is more than simply the sum of the meanings of the individual sentences“ (Grabe und Kaplan 1996, S. 4). Schreiben im Sinne von *Composition* kann man also als strategische Handlung bezeichnen, wobei sich dem kognitiven System der schreibenden Person gleichzeitig folgende Ansprüche stellen:

- *Knowledge*: transforming incoherent thought [...] into a highly conceptualized and precisely related knowledge network;
- *Written speech*: expressing that knowledge network within the linguistic and discourse conventions of written prose; and
- *Rhetorical problem*: conforming to the structures posed by the writer's mental representations of the purpose in writing, his role as a writer and his sense of audience (Flower und Hayes 1981).

Die Transformation von lose organisiertem Wissen in adressatengerechte und überzeugende Texte im fremdsprachlichen *Code* stellt eine Herausforderung dar, welcher aus der Perspektive des Kompetenz*erwerbs* nur durch eine prozessorientierte Schreibdidaktik begegnet werden kann (vgl. im Detail Keller 2013). Lernende müssen die Möglichkeit haben, Modelle von guten Texten im Detail zu analysieren und dabei nicht nur Oberflächenphänomene, sondern auf Tiefenstrukturen der Textorganisation zu erleben und für das eigene Schreiben produktiv zu machen. Ebenso brauchen sie sachgerechte, lerndienliche Rückmeldungen nicht nur zu (Oberflächen-)Fehlern sondern zum *rhetorical problem space sensu* Flower und Hayes (1981). Dazu gehört, dass das Schreiben als soziale Situation erlebbar wird, wobei die Absichten der schreibenden Person, die Erwartungen der Leserinnen und Leser sowie der Kontext des Schreibens aufeinander bezogen werden. Die kompetente Unterstützung solcher Prozesse setzt wiederum hoch entwickeltes fachdidaktisches Wissen auf Seiten der Lehrpersonen voraus (Parr und Timperley 2010).

2.3 Erfassung von Fremdsprachenkompetenzen

International hat es sich mittlerweile bei der Erfassung von Fremdsprachenkompetenzen durchgesetzt, dass sowohl produktive als auch rezeptive kommunikative Kompetenzen standardmäßig erfasst werden. So umfasst die computerbasierte Form des TOEFL einen Leseverstehenstest (60–80 min), einen Hörverstehenstest (60–90 min), einen Test zum Sprechen (20 min) und einen Test zum Schreiben (50 min), der aus zwei Schreibaufgaben besteht. Im deutschen Kontext wurden am Institut zur Qualitätsentwicklung im Bildungswesen (IQB) im Rahmen der Arbeiten zur Evaluation der Bildungsstandards für das Ende der Sekundarstufe I Items zum Lese- und Hörverstehen sowie Aufgaben zum Schreiben entwickelt (vgl. u. a. Rupp und Porsch 2010), die allesamt mit dem GER verlinkt wurden (vgl. Harsch et al. 2010). Bislang wurden aus diesem Pool allerdings nur die Items zum Lese- und Hörverstehen in den Ländervergleichen zur Erreichung der Vorgaben in den KMK-Bildungsstandards eingesetzt (vgl. Stanat et al. 2016), nicht zuletzt aus auswertungsökonomischen Gründen.

Die am IQB entwickelten Schreibaufgaben folgten dabei einem *Uni-Level-Ansatz* (vgl. Porsch und Köller 2010). In diesem Paradigma bearbeitet eine Schülerin bzw. ein Schüler eine Aufgabe, die hinsichtlich der Anforderungen und der nachfolgenden Kodierung bzw. Bewertung auf ein GER-Niveau festgelegt ist. So kann festgestellt werden, ob sie oder er dieses Niveau erreicht, nicht erreicht oder überschreitet. Wenn sie/er das Niveau nicht erreicht hat oder deutlich über diesem liegt, können jedoch keine Aussagen getroffen werden, welchem Niveau sie/er exakt zuzuordnen ist. Deshalb sollte eine Schülerin bzw. ein Schüler mehr als eine Aufgabe bearbeiten, um reliable Aussagen über ihr bzw. sein Kompetenzniveau zu erhalten.

Beim *Multi-Level-Ansatz* werden die Schreibprodukte der Schülerinnen und Schüler allen Stufen des GER durch ein niveaustufenübergreifendes Beurteilungsverfahren zugeordnet. Ein Beispiel für den *Multi-Level-Ansatz* sind die Schreibaufgaben der DESI-Studie, die in den Jahren 2003/4 durchgeführt wurde und deren Beurteilungsskalen die Niveaustufen A1 bis B2+ berücksichtigen (vgl. Harsch 2006; Beck und Klieme 2007; Harsch et al. 2007). Auch die Sprachtests der Cambridge Universität (z. B. ESOL) folgen diesem Ansatz. So werden die gewonnenen Schreibprodukte mit Hilfe einer gemeinsamen Skala mit entsprechenden Deskriptoren den Niveaustufen A2 bis C2 zugeordnet (Hawkey und Barker 2004). Das Gleiche gilt für den TOEFL, für den sechs Stufen (0–5) definiert sind, die in der Vergangenheit auch schon mit den GER-Stufen verlinkt wurden (vgl. Tannenbaum und Wylie 2005).

Porsch und Köller (2010) zeigen in einer empirischen Studie die hohen Korrelationen zwischen *Uni-Level-* und *Multi-Level-Aufgaben*. Hinsichtlich der Validität unterschieden sich beide Ansätze in der Untersuchung auch nicht. Für die vorliegende Untersuchung wurde ein *Multi-Level-Ansatz* verfolgt, der im folgenden Abschnitt beschrieben wird.

2.3.1 Erfassung von fremdsprachlichen Schreibkompetenzen in der Oberstufe

Aus der Konzeptualisierung des fremdsprachlichen Schreibens in den Bildungsstandards für die gymnasiale Oberstufe ergibt sich die Herausforderung, Aufgabenformate zu identifizieren oder zu entwickeln, mit welchen sich diese komplexen Funktionen des Schreibens unter den Bedingungen einer Large-Scale-Testung reliabel, valide und objektiv messen lassen. Eine Analyse von unterschiedlichen Formaten von Schreibtests, welche im Rahmen einer Vorstudie zur hier geschilderten Untersuchung durchgeführt wurde, ergab, dass sich die curricularen Vorgaben zum fremdsprachlichen Schreiben in der gymnasialen Oberstufe mit dem TOEFL *Internet Based Test* (IBT) curricular valide abbilden lassen (Fleckenstein et al. im Druck). Dieser Test wurde vom *Educational Testing Service* (ETS) in Princeton entwickelt, um die Sprachfertigkeiten von internationalen Studienplatzbewerberinnen und -bewerbern in den USA und Kanada zu überprüfen, deren Muttersprache nicht Englisch ist (zu Details des Tests s. www.test.org/toefl). Mittlerweile wird er aber weltweit eingesetzt, um die Englischkompetenzen von Fremdsprachenlernenden zu erfassen.

Im Schreibtest des TOEFL werden zwei Typen von Aufgaben (*Tasks*) verwendet, welche komplementär verwendet werden und je unterschiedliche Teildimensionen des Gesamtkonzepts „L2-Schreiben“ operationalisieren. Die *Independent Tasks* erfordern vom Lernenden, in exakt 30 min einen argumentativen Text zu schreiben, in

dem einer vorgegebenen Aussage zugestimmt/nicht zugestimmt wird (z. B. Aufgabe *Teachers*: „Die Fähigkeiten von Lehrkräften, ein gutes Verhältnis zu den Schülerinnen und Schülern zu haben, ist wichtiger als ein hohes Wissen im Fach, das man unterrichtet."). Dazu sollen Argumente benannt werden, warum der Aussage zugestimmt/nicht zugestimmt wird, und der Aufsatz mit Einleitung und Schluss angemessen gerahmt werden. Dieser Aufgabentyp operationalisiert jene Funktion des Schreibens, die oben als „argumentativ" bezeichnet wurde. Solche *Independent Tasks* werden hoch bewertet, wenn sie gut organisiert sind und spezifische Evidenz zur Unterstützung der verwendeten Argumentationskette enthalten. Zudem wird erwartet, dass die englische Sprache akkurat und textsortengerecht verwendet wird, um die persönliche Mitteilungsabsicht der Schreibenden auszudrücken (Rupp et al. 2019).

Der zweite Typus von Schreibaufgaben ist die *Integrated Task*. Hier wird von den Schülerinnen und Schülern zunächst verlangt, dass sie einen Text (Umfang 250–320 Wörter) zu einem wissenschaftlichen Thema lesen (ca. 5 min; z. B. einen Text über elektronische Wahlsysteme in den USA). Anschließend erhalten Sie einen zwei- bis dreiminütigen auditiven Stimulus, in dem eine oppositionelle Rolle zum gelesenen Text vorgestellt wird. Die Schülerinnen und Schüler sollen dann in exakt 20 min schriftlich herausarbeiten (150–225 Wörter), inwiefern sich die Aussagen beider Quellen widersprechen. Diese Schreibaufgabe verlangt also das Verfassen von *Synthesis Texts* (van Ockenburg et al. 2016), wobei eine Kombination verschiedener Sprachkenntnisse und Strategien erforderlich ist: Die Lernenden müssen das Inputmaterial zunächst grob verstehen und danach vertieft analysieren, um geeignetes Inputmaterial für den eigenen Text zu generieren. Dieses muss anschließend sachlogisch organisiert und sinnhaft versprachlicht werden. Von den Lernenden wird jedoch nicht erwartet, dass sie ihre eigene Meinung zu dieser Aufgabe äußern, vielmehr geht es darum, Hauptpunkte einer wissenschaftlichen Argumentation zu erkennen und einander gegenüberzustellen (Plakans 2010). *Integrated Tasks* werden hoch bewertet, wenn sie die zentralen Argumente der Inputtexte klar erfassen, sich widersprechende Argumente im schriftlichen und gesprochenen Text kontrastieren und in einer klaren, konzisen Sprache verfasst sind.

Die curriculare Validität des TOEFL IBT ist zwar nicht in dem Sinne gegeben, dass die Lehrpersonen konkret im Unterricht in deutschen Schulen diese Aufgaben einsetzen würden – besonders *Integrated Tasks* gehören nicht zum typischen Unterrichtsrepertoire von Gymnasiallehrkräften in der Oberstufe (Keller 2013). Die darin enthaltenen Arten von Schreibaufgaben bilden jedoch zentrale Konzepte des fremdsprachlichen Schreibens, welche in den übergeordneten Bildungsstandards ausgedrückt sind, auf ökonomische und valide Weise ab. Zudem hat Fleckenstein (2017) gezeigt, dass sich die im TOEFL IBT getesteten Fremdsprachkompetenzen gut in der Systematik des GER ausdrücken lassen. Es besteht also eine Kongruenz zwischen dem GER und den Bildungsstandards für das Schreiben in der Oberstufe auf der einen, und den Schreibaufgaben des TOEFL IBT Tests auf der anderen Seite, welche Voraussetzung ist für die in dieser Studie verfolgten Fragestellungen.

2.4 Profiloberstufe

In Schleswig-Holstein werden die Schülerinnen und Schüler wie in vielen anderen Ländern (u. a. Baden-Württemberg und Hamburg) in allgemeinbildenden Gymnasien nicht mehr in Grund- und Leistungskursen unterrichtet, sondern im Klassenverband in Profiloberstufen. Dort belegen sie Mathematik, Deutsch, eine Fremdsprache (in der großen Mehrzahl Englisch) und ein profilgebendes Fach auf erhöhtem Anforderungsniveau (vierstündig). Fünf Profile (naturwissenschaftlich, sprachlich, gesellschaftswissenschaftlich, sportwissenschaftlich, ästhetisch) legen darüber hinaus Schwerpunkte fest.[1] Im sprachlichen Profil werden in der Qualifizierungsphase typischerweise drei Fremdsprachen, eine oder zwei vierstündig, eine oder zwei dreistündig unterrichtet. Dies mündet in bis zu 11 Stunden Fremdsprachenunterricht.

Neben Deutsch und der ersten bzw. fortgeführten Fremdsprache (zwei der drei Kernfächer; acht Wochenstunden) muss dagegen im naturwissenschaftlichen Profil keine weitere Fremdsprache in der Qualifizierungsphase belegt werden. Dies gilt ebenfalls für die anderen drei nicht-sprachlichen Profile. Hinsichtlich der Zusammensetzung der Schülerschaft in den unterschiedlichen Profilen haben früheren Untersuchungen gezeigt, dass die kognitiven Grundfähigkeiten am höchsten im naturwissenschaftlichen Profil ausgeprägt sind (vgl. Köller 2016) und dort auch deutlich mehr Schüler als Schülerinnen zu finden sind. Im sprachlichen Profil dominieren Schülerinnen ebenso wie im ästhetischen. Dagegen zeigt sich auch ein Überschuss an Schülern im gesellschaftswissenschaftlichen und sportwissenschaftlichen Profil (Leucht und Köller 2016).

Leucht et al. (2015) haben rezeptive Englischleistungen im Laufe der gymnasialen Oberstufe in Schleswig-Holstein untersucht. Berücksichtig wurden die Ergebisse von über 1000 Schülerinnen und Schülern im Lese- und Hörverstehen, die Mitte der 11. Jahrgangsstufe (T1) und Ende der 13. Jahrgangsstufe (T2) im neunjährigen Gymnasium gewonnen wurden. Die Studie erlaubt sowohl die Abschätzung von Kompetenzgewinnen in der gymnasialen Oberstufe als auch die nach Profilen aufgeteilte Schätzung von Kompetenzzuwächsen. Deskriptiv zeigte sich zunächst, dass die Lesekompetenzen im Fach Englisch über 2,5 Schuljahre nur gering stiegen ($d = 0{,}25$), die Kompetenzen im Bereich des Hörverstehens aber deutlich ($d = 1{,}18$). Profilspezifische Auswertungen ergaben, dass die Schülerinnen und Schüler des sprachlichen Profils die höchsten Leistungen zu beiden Zeitpunkten aufwiesen und auch die Zuwächse von T1 nach T2 im sprachlichen Profil am höchsten ausfielen, und dies, obwohl in allen Profilen Englisch vierstündig auf erhöhtem Anforderungsniveau unterrichtet wurde. Die zweithöchsten Kompetenzstände und Leistungszuwächse wurden im naturwissenschaftlichen Profil erreicht. Die übrigen Profile lagen in ihren Leistungen und Zuwächsen dicht beieinander. Leucht et al. bieten Erklärungsansätze für die günstigen Leistungsverläufe im sprachlichen Profil an: (1) Transfereffekte der anderen in der Oberstufe belegten modernen Fremdsprachen; (2) höhere Unterrichtsqualität durch höhere professionelle Kompetenzen der Englischlehrkräfte

[1] Die Landesregierung in Schleswig-Holstein hat inzwischen angekündigt, dass die Profiloberstufe ab dem Schuljahr 2021/22 verpflichtend leicht modifiziert wird. Für die Fragestellungen dieser Studie sind die angekündigten Reformen allerdings irrelevant.

im sprachlichen Profil; (3) verstärkte, interessengeleitete Freizeitaktivitäten in der Fremdsprache Englisch.

2.5 Fragestellungen und Hypothesen

Mit der vorliegenden Untersuchung sollten insgesamt vier Fragestellungen bearbeitet werden. In einem ersten Schritt wurde der Frage nachgegangen, ob sich die Schreibkompetenzen von Schülerinnen und Schülern im Fach Englisch in der gymnasialen Oberstufe psychometrisch zufriedenstellend (hinreichende Reliabilität und Validität) mit dem gewählten Paradigma (*Multi-Level-Aufgaben* aus dem TOEFL) erfassen lassen. In einem zweiten Schritt sollte ergänzend untersucht werden, ob sich mit dem gewählten Paradigma auch Kompetenzzuwächse abbilden lassen. In einem dritten Schritt sollten die Befunde von Leucht et al. (2015) zu Profilunterschieden in den rezeptiven Englischleistungen und ihrer Veränderung (Lesen und Hören) repliziert werden. Dazu wurde vorhergesagt, dass Schülerinnen und Schüler, die ein fremdsprachliches Profil in der gymnasialen Oberstufe besuchen, höhere rezeptive Leistungen zeigen und auch ihre Leistungszuwächse über ein Jahr höher ausfallen sollten als die der Schülerinnen und Schüler in den anderen Profilen. Weiterhin wurde vorhergesagt, dass solch ein Ergebnismuster auch für sprachproduktive Kompetenzen nachweisbar sein sollte. In einem letzten Schritt wurde der Frage nachgegangen, welche Leistungsstände Schülerinnen und Schüler unterschiedlicher Profile mit Bezug auf die in den Bildungsstandards der KMK festgelegten Ziele im Bereich der Schreibkompetenzen (GER-Niveau B2 und höher) erreichen. Um Unterschiede in der Komposition der Profile in den Analysen kontrollieren zu können, wurden das Geschlecht und die kognitiven Grundfähigkeiten als Kovariaten berücksichtigt. Frühere Arbeiten zu Profiloberstufen (vgl. Leucht et al. 2015) hatten gezeigt, dass Mädchen bzw. Frauen gehäuft in sprachlichen und deutlich seltener in naturwissenschaftlichen Profilen vertreten sind. Auch hatte sich dort gezeigt, dass sich die Schülerinnen und Schüler je nach Profil in ihren kognitiven Grundfähigkeiten unterscheiden, d. h. Schülerinnen und Schüler in naturwissenschaftlichen Profilklassen wiesen signifikant höhere Mittelwerte auf als in den anderen Profilen.

3 Methode

Die Daten der vorliegenden Arbeit stammen aus dem Projekt *Measuring English Writing Skills at Secondary Level (MEWS)*, das als Kooperationsprojekt zwischen dem Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN) und der Fachhochschule Nordwestschweiz (FHNW), Pädagogische Hochschule, Basel durchgeführt wird. MEWS wird von der Deutschen Forschungsgemeinschaft (DFG) und vom Schweizer Nationalfonds (SNF) gefördert und untersucht in einem mehrebenenanalytischen Messwiederholungsdesign Englischleistungen in der Qualifikationsphase der gymnasialen Oberstufe (11. Jahrgangsstufe im achtjährigen Gymnasium) in deutschen und Schweizer Oberstufen. Für die vorliegende Untersuchung, in der es um die Erreichung von Zielvorgaben der KMK bezüglich

sprachproduktiver Kompetenzen geht, wurde eine Beschränkung auf die deutschen Daten vorgenommen.

3.1 Datenerhebung

Die Untersuchung fand im Schuljahr 2016/17 statt, der erste Messzeitpunkt (T1) war im September 2016, der zweite (T2) im Juni 2017. Die Datenerhebung wurde von der IEA Hamburg übernommen. Die IEA Hamburg ist in Deutschland für die Datenerhebungen aller Large-Scale-Assessments verantwortlich (u. a. IQB-Bildungstrend, TIMSS, PIRLS und PISA) und verfügt über breite Erfahrungen in der Testung von Schülerstichproben. Vorab wurden Grundinformationen (u. a. Geschlecht, Oberstufenprofil, Noten im letzten Zeugnis in Deutsch, Englisch und Mathematik) zu den Schülerinnen und Schülern mit Schülerteilnahmelisten (STL) erhoben. Die STL wurden den Schulen zugesendet und jeweils von Schulkoordinatorinnen/-koordinatoren ausgefüllt.

Die Testsitzungen in den Schulen wurden von trainierten Testleiterinnen und -leitern durchgeführt. Zum Einsatz kam ein Computer-basiertes Test- und Befragungssystem, das am IPN entwickelt wurde. Sämtliche Testaufgaben und Fragebögen wurden dementsprechend an Notebooks bearbeitet. Jede Testsitzung (zu T1 und T2) begann mit allgemeinen Informationen zur Datenerhebung. Es folgten zwei Schreibaufgaben (insgesamt 60 min), anschließend eine Pause, dann 60 min lang Tests zum Lese- und Hörverstehen. Nach einer weiteren Pause wurden zwei kurze Tests zur Feststellung der kognitiven Grundfähigkeiten bearbeitet (nur zu T1), anschließend wurden direkt die Fragebogenitems administriert. Die Gesamtdauer der Datenerhebung umfasste so rund dreieinhalb Zeitstunden. Die Studie mit all ihren Instrumenten und Auflagen zum Datenschutz war zuvor vom zuständigen Ministerium in Schleswig-Holstein genehmigt worden. Die Testung und Befragung waren für alle Schülerinnen und Schüler freiwillig.

3.2 Stichprobe

Die Stichprobe bestand aus insgesamt $N = 990$ Schülerinnen (57,1 %) und Schülern, die zu Beginn der Qualifikationsphase (11. Jahrgangsstufe) an Schulen in Schleswig-Holstein gezogen wurden (mittleres Alter $M = 16{,}9$ Jahre, $SD = 0{,}56$). Die Schülerinnen und Schüler stammten aus 37 allgemeinbildenden Gymnasien (von 99 in Schleswig-Holstein). Der Anteil der Jugendlichen mit Migrationshintergrund war sehr gering, 97 % der Schülerinnen und Schüler waren in Deutschland geboren, in knapp 10 % der Fälle war mindestens ein Elternteil im Ausland geboren. In der Stichprobe besuchten 35,4 % ein naturwissenschaftliches Profil, 22,7 % ein sprachliches, 32,5 % ein gesellschaftliches, 4,1 % ein ästhetisches und 5,3 % ein sportwissenschaftliches Profil. Wegen der kleinen Fallzahlen (39 und 50) wurden das sportwissenschaftliche und das ästhetische Profil zu einer Gruppe zusammengefasst.

Von den ursprünglich gezogenen $N = 990$ Schülerinnen und Schülern nahmen an der ersten Testung $N_{T1} = 838$ teil (TN-Gruppe; Ausschöpfungsquote 84,6 %). Da von ihnen und den nicht teilnehmenden Jugendlichen (NTN-Gruppe) Noten des letzten Zeugnisses (Ende 10. Jahrgangsstufe) in den drei Fächern Deutsch,

Mathematik und Englisch auf den STL vorlagen, konnten Selektivitätsanalysen durchgeführt werden. Die Mittelwertvergleiche (Signifikanzniveau $\alpha = 0{,}05$) zeigten durchgängige keine signifikanten Unterschiede zwischen beiden Gruppen (Deutschnote: $M_{TN} = 8{,}50$, $SD_{TN} = 2{,}68$; $M_{NTN} = 8{,}28$, $SD_{NTN} = 2{,}64$; $t_{936} = 0{,}85$, $p(2s) = 0{,}397$; Englischnote: $M_{TN} = 8{,}94$, $SD_{TN} = 2{,}71$; $M_{NTN} = 8{,}45$, $SD_{NTN} = 2{,}57$; $t_{938} = 1{,}88$, $p(2s) = 0{,}060$; Mathematiknote: $M_{TN} = 8{,}50$, $SD_{TN} = 3{,}03$; $M_{NTN} = 8{,}32$, $SD_{NTN} = 3{,}05$; $t_{936} = 0{,}62$, $p(2s) = 0{,}536$). Damit ergaben sich keine Hinweise auf eine substanzielle Verzerrung in der Stichprobe aufgrund der Nichtteilnahme. Entsprechend wurde festgelegt, dass die $N_{T1} = 838$ Schülerinnen und Schüler die Analysestichprobe darstellen sollten. Von diesen nahmen $N_{T2} = 654$ Schülerinnen und Schüler an T2 teil. Für Schülerinnen und Schüler, die nicht an T2 teilnahmen, wurde entschieden, mit Hilfe der Informationen von T1 die Testleistungen zu T2 zu schätzen (vgl. hierzu Abschnitt 3.3). Dementsprechend basieren auch die Analysen, die beide Erhebungszeitpunkte umfassen, auf Daten von 838 Schülerinnen und Schülern.

3.3 Leistungstests und Skalierung

In der vorliegenden Untersuchung wurden rezeptive (Lese- und Hörverstehen) und produktive (Schreiben) Kompetenzen in der Fremdsprache Englisch getestet. Darüber hinaus wurden zwei Untertests aus dem Kognitiven Fähigkeitstest für 4. bis 12. Klassen, Revision (KFT 4–12 + R; Heller und Perleth 2000) eingesetzt.

Hör- und Leseverstehen Englisch Zur Erfassung des Lese- und Hörverstehens wurde eine Teilmenge der Aufgaben des Ländervergleichs von Köller et al. (2010) eingesetzt. Die Aufgaben erfassen die in den Bildungsstandards für die erste Fremdsprache Englisch (KMK 2004) beschriebenen Anforderungen zum Lese- und Hörverstehen. Insgesamt wurden 133 Lese- und 118 Hörverstehenitems eingesetzt. Auf einen Lese- oder Hörtext (Aufgabenstamm) folgten drei bis acht Items, die gemeinsam mit dem Stamm ein Testlet konstituierten. Für die Testlets lagen aus früheren Studien (u. a. Köller et al. 2010) Bearbeitungszeiten vor, so dass mehrere Testlets jeweils zu einem Block von 15 min zusammengefasst wurden. Jeder Schülerin und jedem Schüler wurden zwei solcher 15-Minuten-Blöcke zum Leseverstehen und zwei zum Hörverstehen zur Bearbeitung vorgegeben. Die Reihenfolge der Blöcke und der zuerst erhobenen Kompetenz (Lesen vs. Hören) wurden in einem so genannten Multi-Matrix-Design variiert. Dadurch wurde sichergestellt, dass später geschätzte Itemparameter unabhängig von ihrer Position im Testheft waren. Die Schätzung von Itemparametern erfolgte auf Basis längsschnittlicher mehrdimensionaler zweiparametriger Item-Response-Modelle in M*plus* (Version 8; Muthén und Muthén 1998–2017). Als Schätzer für die Personenfähigkeiten wurden Plausible Values (PVs; 15 pro Person) bestimmt. In die Schätzung der PVs gingen neben den Lösungsvektoren in einem Hintergrundmodell sämtliche Informationen aus der STL und dem Schülerfragebogen nach vorheriger Hauptkomponentenanalyse ein. Analog zum Vorgehen des IQB wurde die Testletstruktur bei den Analysen ignoriert. Durch die längsschnittliche PV-Ziehung konnten auch Fähigkeitsschätzungen zu T2 für die Schülerinnen und Schüler vorgenommen werden, die nicht mehr an der

zweiten Testung teilgenommen hatten. Zur besseren Interpretierbarkeit der PVs wurden diese entsprechend dem Vorgehen bei internationalen Vergleichsstudien (z. B. PISA, vgl. Reiss et al. 2016) auf eine Metrik mit einem Mittelwert von $M = 500$ und einer Standardabweichung von $SD = 100$ zu T1 transformiert. Für die Werte zu T2 bedeutete diese Transformation, dass die Mittelwertsdifferenz zu T1 direkt als Kompetenzzuwachs interpretiert werden konnte. Die Reliabilitäten der PVs lagen für die Lesekompetenzen bei 0,92 (T1) und 0,76 (T2), für das Hörverstehen ergaben sich PV-Reliabilitäten von 0,85 (T1) und 0,72 (T2).

Schreibaufgaben Die hier verwendeten Schreibaufgaben stammen aus dem TOEFL *Internet Based Test* (IBT, s. oben). Insgesamt wurden zwei *Independent Tasks* (*Teachers* und *TV Advertising*) und zwei *Integrated Tasks* (*The Chevalier* und *Voting Machines*) verwendet (zur genaueren Beschreibung der Aufgaben vgl. Rupp et al. 2019 und Anhang, Tab. 7). Jede Schülerin/jeder Schüler musste zu jedem Messzeitpunkt eine *Integrated* und eine *Independent Task* bearbeiten. Dabei wurde darauf geachtet, dass keine Aufgabe wiederholt vorgegeben wurde. Hatten die Schülerinnen und Schüler beispielsweise zu T1 *Teachers* und *The Chevalier* bearbeitet, erhielten sie zu T2 *TV Advertising* und *Voting Machines*. Die Kodierung der Texte erfolgte sowohl durch geschulte Kodierinnen/Kodierer (Human Raters; HR) als auch mittels computerbasierter (maschineller) Kodierung. Das genaue Vorgehen ist bei Rupp et al. (2019) dargestellt, weshalb die Beschreibung im Folgenden relativ kurzgehalten wird. Jedes Essay wurde von zwei unabhängigen HR auf Basis der Deskriptoren bei der Bewertung von TOEFL-Essays auf einer holistischen Skala von 0 bis 5 Punkten kodiert. Die Beschreibung der sechs Stufen (Punkte 0 bis 5) findet sich für beide Aufgabentypen bei ETS und ist direkt im Internet abrufbar (www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf; Abruf 03.05.2019). Je höher der Wert, desto höher die Textqualität (vgl. hierzu auch Abschnitt 2.3.1). Die Interrater-Übereinstimmung (quadriertes gewichtetes Kappa) variierte in unserer Studie je nach Aufgabe und Messzeitpunkt zwischen 0,568 (T2, *TV Advertising*) und 0,874 (T1, *The Chevalier*) und kann als sehr zufriedenstellend bezeichnet werden.

Über die HR hinaus wurden alle Texte automatisch durch eine Computer-Software kodiert (*Automated Essay Evaluation*; AEE). Beim AEE werden grundlegende, maschinell gut erfassbare Textmerkmale durch *Natural Language Processing* (NLP) elektronisch mit der Software E-rater® ausgewertet (Ramineni et al. 2012). E-rater® basiert auf den Arbeiten von Burstein et al. (1998) und wurde seither bei ETS laufend weiterentwickelt und evaluiert (Ramineni et al. 2012). Wichtige Kategorien, so genannte *Macro-Features*, die E-rater® erfasst, sind Grammatik, Sprachgebrauch, Sprachmechanik, Stil, Organisation und Textaufbau. Zusätzlich erfasst E-rater® noch Textqualitäten (*Positive Features*) wie Gebrauch von Pronomina und Wortkollokationen, sowie zwei Maße der lexikalischen Komplexität. Sämtliche Textmerkmale von E-rater® werden mittels regressionsanalytischer Verfahren oder Techniken des Maschinenlernens mit den Bewertungen der HR in Beziehung gesetzt und erhalten so Vorhersagegewichte für die Werte der HR. Typischerweise wird die Vorschrift, mit der die Features die HR vorhersagen, an der halben Stichprobe der Texte ermittelt und an der anderen Hälfte validiert. In der vorliegenden Untersuchung wurden sowohl generische Modelle (Gewichte der Macro-Features waren für die zwei un-

terschiedlichen Essays jedes *Tasks* identisch) als auch promptspezifische Modelle bestimmt, bei denen für jedes der vier Essays ein Vorhersagemodell generiert wurde (zu den Details des Vorgehens vgl. Rupp et al. 2019). Dabei erwiesen sich promptspezifische Modelle des Maschinenlernens auf der Basis der Texte von beiden Messzeitpunkten als die zuverlässigsten. Sie lieferten Bewertungen der Texte, die zwischen 0,5 und 5,5 schwankten und im Gegensatz zu den HR auch Dezimalwerte annehmen konnten. Gemittelt über beide Messzeitpunkte ergaben sich sehr hohe Übereinstimmungen zwischen Maschinenratings und den HR, die zwischen 0,784 (*Voting Machines*) und 0,852 (*The Chevalier*) lagen.

Wie im Theorieteil beschrieben, besteht die Herausforderung einer solchen Studie darin, Schreibleistungen im Sinne von *Composition* zu konzeptualisieren und zu beurteilen. In den menschlichen Urteilen ist dies gegeben, da die holistischen *Scoring rubrics* des TOEFL IBT explizit Beschreibungen wie *effectively addressing a topic* oder *displaying unity of progression* enthalten (vgl. ets.org). Allerdings analysieren und verarbeiten automatisierte Beurteilungssysteme wie E-rater® Texte grundsätzlich anders als Menschen. Während Menschen sich direkt einen Gesamteindruck eines Texts bilden können, welcher auf Urteilen zu Diktion, Flüssigkeit oder Argumentation aufbaut, verwenden Maschinen einzelne Approximationen oder Korrelate dieser Elemente, aus welchen mittels Maschinenlernen die Textqualität vorausgesagt werden kann (vgl. dazu im Detail Deane 2013). Auch neueste automatisierte Beurteilungssysteme sind nicht in der Lage, menschliche Fähigkeiten wie jene zur Konzeptualisierung oder Darstellung überzeugender Argumente valide einzuschätzen. Sie erfassen deshalb notwendigerweise einen engeren Kompetenzbereich als menschliche Beurteilerinnen/Beurteiler, wobei die erfassten Textmerkmale sich primär auf der Ebene von Sätzen und Phrasen befinden (Deane 2013).

Es gibt allerdings aus kognitiver Sicht eine vertiefte Verbindung zwischen den maschinell messbaren linguistischen Kompetenzen und den Fähigkeiten von Schreibenden, argumentativ komplexe und gehaltvolle Texte in der Fremdsprache zu verfassen. Schreibende, welche solche linguistischen Fähigkeiten nicht ausreichend erworben oder automatisiert haben, verfügen auch nicht über die kognitiven Ressourcen, die Voraussetzung sind für höhere Schreibfunktionen wie Argumentation, Leserführung oder Herstellung von inhaltlicher Kohärenz (McCutchen 2000). Mit anderen Worten, jene Lernenden, welche hohe Fähigkeiten im Bereich Vokabular, Grammatik und Orthographie mitbringen, sind auch jene, welche raffinierte Argumentationen herstellen und überzeugend formulieren können.

Zudem lagen in dieser Studie von jedem Text zwei menschliche Urteile vor, so dass die menschlichen Urteile gegenüber der maschinellen das doppelte Gewicht hatten. Diese zwei HR sowie das Maschinenrating gingen dann in die gemeinsame Skalierung der Schreibleistung ein.

Zur Berechnung von Personenfähigkeiten wurden wie beim Lese- und Hörverstehen pro Messzeitpunkt 15 PVs aus zweidimensionalen IRT-Modellen mit Hintergrundmodell gezogen. Die PVs erreichten Reliabilitäten von 0,94 (T1) bzw. 0,85 (T2). Wiederum wurde eine Standardisierung der PVs auf einer Skala mit $M = 500$ und $SD = 100$ zu T1 vorgenommen.

Kognitive Grundfähigkeiten Die beiden Skalen Figurenanalogien (25 Items) und Wortanalogien (20 Items) aus dem KFT 4–12 + R (Heller und Perleth 2000) dienten zu T1 zur Erfassung der kognitiven Grundfähigkeiten. Für beiden Skalen wurden je 15 PVs aus einem zweidimensionalen Modell gezogen. Für die vorliegende Untersuchung wurden beide Teilskalen zu einem Gesamtwert (insgesamt 15 entsprechend der Zahl der PVs) zusammengefasst, der eine Reliabilität von 0,86 erreichte. Auch hier wurde die 500/100-Standardisierung vorgenommen.

Zeugnisnoten Zeugnisnoten für jede Schülerin und jeden Schüler (Ende der 10. Jahrgangsstufe) wurden der STL entnommen (s. oben). Dort waren die Deutsch-, Englisch- und Mathematiknote (0 bis 15 Punkte) berichtet.

3.4 Statistische Analysen

Alle statistischen Analysen wurden mit dem Softwarepaket M*plus* (Version 8, Muthén und Muthén 1998–2017) durchgeführt. Der hierarchische Charakter der Daten (Schülerinnen und Schüler innerhalb von Klassen bzw. Schulen) bedingt, dass die Individualstichprobe keine unabhängige Stichprobe ist, Schülerinnen und Schüler sich innerhalb von Schulen vielmehr ähneln, dafür aber erhebliche Unterschiede zwischen Klassen bzw. Schulen auftauchen. In der Tat zeigten sich substanzielle Differenzen in den Leistungsmaßen zwischen den Schulen. Die Intra-Klassen-Korrelationen (ICCs) schwankten bei den Englischtests zwischen 0,084 (T1 Hören) und 0,146 (T2 Schreiben). Solche Klumpenstichproben führen bei der Schätzung von statistischen Parametern dazu, dass die dazugehörigen Standardfehler verzerrt sind und damit die statistische Inferenz invalide wird. Die Nutzung der Option Type = Complex in M*plus* erlaubt die Korrektur und so Ermittlung unverzerrter Standardfehler (vgl. Muthén und Satorra, 1995). Alle Analysen der vorliegenden Untersuchung wurden dementsprechend mit dieser Option durchgeführt.

Weiterhin lagen für jede Leistungsvariable 15 PVs vor. Als Folge wurde jede der unten präsentierten Analysen fünfzehnmal durchgeführt und die Befunde entsprechend dem Vorgehen bei Rubin (1987) gemittelt.

Missing data stellten kein weiteres Problem dar, da für alle Leistungsmaße PVs generiert wurden, auch wenn die Tests zu T2 nicht mehr bearbeitet worden waren (s. oben). Bei den Fachnoten fehlten die Daten von lediglich 25 Schülerinnen und Schülern (3 %), diese wurden mittels *Single Imputation* geschätzt.

3.5 Definition von GER-Stufen

Um die Fragestellung zur Erreichung normativer Vorgaben der Bildungsstandards im Bereich Schreiben bearbeiten zu können, war zusätzlich ein *Standard-Setting* (Cizek 2012) erforderlich. Solche *Standard-Settings* erlauben es, für Testergebnisse bzw. einen Test Kompetenzstufen zu definieren, mit deren Hilfe die Testteilnehmerinnen und -teilnehmer kriterial hinsichtlich ihrer erreichten Leistungsstände beschrieben werden können. Konkret wurde in der vorliegenden Studie angestrebt, Schreibleistungen, die sowohl auf der TOEFL-Metrik (0–5) als auch nach Transformation der IRT-skalierten Leistungswerte auf einer 500/100-Metrik vorlagen, den GER-Stufen

zuzuordnen. Details zum *Standard-Setting* der Daten sind bei Fleckenstein et al. (im Druck) beschrieben, weshalb das Vorgehen im Folgenden stark verkürzt erläutert wird. Zur Anwendung kam die sogenannte *Performanz-Profil-Methode* (Hambleton et al. 2000). Dabei wurden ausgewählte geschriebene Texte (je eine *Integrated* und eine *Independent Task*) von Schülerinnen und Schülern, die in ihren Schreibkompetenzen breit streuten, einem *Panel* von 12 fachdidaktischen und psychometrischen Expertinnen und Experten zur Definition von *Cut-Scores* vorgelegt. Die *Cut-Scores* definieren die Grenzen zwischen Niveaus des GER. Zur Festlegung der Grenze zwischen den Niveaus A2 und B1 wurden die Expertinnen und Experten beispielsweise gebeten, ein Profil von zwei Texten einer Schülerin bzw. eines Schülers herauszusuchen, die/der soeben das Niveau B1 erreicht. Diese Wahl des Profils musste konsensuell erfolgen. Anschließend wurde ermittelt, welchen TOEFL-Score diese Schülerin/dieser Schüler gemittelt über zwei Texte und drei Ratings (Human/Human/E-rater®) erreicht hatte. Diese Festlegung erbrachte folgende *Cut-Scores* auf der TOEFL-Metrik (0–5). Grenze A2/B1: 2,25 Punkte; Grenze B1/B2: 2,75 Punkte; Grenze B2/C1: 3,75 Punkte; und Grenze C1/C2: 4,75 Punkte. Auf der hier gewählten IRT-Metrik ($M = 500$; $SD = 100$) ergaben sich folgende Grenzen: A2/B1: 425 Punkte; B1/B2: 493 Punkte; B2/C1: 630 Punkte; C1/C2: 768 Punkte. Auf die Festlegung einer Grenze zwischen den Niveaus A1 und A2 wurde verzichtet, da keine Schülerinnen und Schüler auf dem niedrigen A1-Niveau erwartet wurden.

4 Ergebnisse

Im ersten Schritt werden Befunde aus Validierungsanalysen der Schreibaufgaben vorgestellt. Es folgen dann deskriptive Befunde zu Leistungsunterschieden zwischen den Profilen und zu Leistungsveränderungen über die Zeit, die durch multivariate Analysen unter Berücksichtigung von Kovariaten inferenzstatistisch abgesichert werden. Schließlich findet eine Verortung der Schreibleistungen auf den Stufen des GER statt.

4.1 Analysen zur Validität der Leistungswerte in den Schreibaufgaben

Um einen ersten Eindruck hinsichtlich der Validität der Schreibleistungen zu erhalten, wurden die PVs zu T1 und T2 mit den Noten in den Fächern Englisch, Deutsch und Mathematik sowie den Lese- und Hörleistungen korreliert. Die Tab. 1 zeigt die entsprechenden Koeffizienten. Erwartungskonform ergaben sich hohe Korrelationen zwischen den drei Englischleistungstests, die zwischen $r = 0{,}42$ und $r = 0{,}60$ schwanken. Die höchste Korrelation ergab sich für die Stabilität der Schreibleistungen ($r = 0{,}74$), die Stabilitäten beider rezeptiver Kompetenzen lagen um $r = 0{,}50$. Hinsichtlich der Korrelationen der Leistungstests mit den Fachnoten am Ende der 10. Jahrgangsstufe ergeben sich plausible Befunde. Die höchsten Zusammenhänge zeigten sich mit der Englischnote (konvergente Validität), die geringsten mit der Mathematiknote (diskriminante Validität). Bemerkenswert ist, dass die Schreibleitungen zu beiden Zeitpunkten Korrelationen mit der Englischnote von über 0,40 aufweisen. Schließlich zeigte sich für die Zusammenhänge der Leistungstests mit dem Maß

Tab. 1 Korrelationen der berücksichtigten Variablen (gemittelt über 15 Datensätze)

	T1 Les	T1 Hör	T1 Schr	T2 Les	T2 Hör	T2 Schr	Kog. Gr	Note-E	Note-D	Note-M
T1 Les	1,00	–	–	–	–	–	–	–	–	–
T1 Hör	0,522	1,00	–	–	–	–	–	–	–	–
T1 Schr	0,487	0,585	1,00	–	–	–	–	–	–	–
T2 Les	0,503	0,418	0,504	1,00	–	–	–	–	–	–
T2 Hör	0,467	0,521	0,597	0,490	1,00	–	–	–	–	–
T2 Schr	0,457	0,549	0,742	0,478	0,535	1,00	–	–	–	–
Kog. Gr	0,383	0,398	0,273	0,288	0,283	0,279	1,00	–	–	–
Note-E	0,213	0,318	0,433	0,231	0,353	0,408	0,116	1,00	–	–
Note-D	0,216	0,260	0,309	0,206	0,270	0,319	0,170	0,611	1,00	–
Note-M	0,122	0,146	0,154	0,156	0,180	0,200	0,183	0,439	0,551	1,00

T1 Beginn der 11. Jahrgangsstufe, *T2* Ende der 11. Jahrgangsstufe, *Les.* Leseverstehen, *Hör.* Hörverstehen, *Schr.* Schreibkompetenz, *Kog. Gr.* Kognitive Grundfähigkeiten, *Note-E* Englischnote im Endzeugnis der 10. Jahrgangsstufe, *Note-D* Deutschnote im Endzeugnis der 10. Jahrgangsstufe, *Note-M* Mathematiknote im Endzeugnis der 10. Jahrgangsstufe; alle Korrelationen weichen signifikant von Null ab ($p < 0{,}01$)

zu den kognitiven Grundfähigkeiten, dass diese durchgängig positiv ausfielen, aber vom Betrag her niedriger als die Korrelationen der Leistungstests untereinander. Insgesamt ergibt sich demnach ein Ergebnismuster in Tab. 1, das die zufriedenstellende Validität des eingesetzten Paradigmas zur Erfassung von Schreibkompetenzen belegt.

4.2 Unterschiede zwischen den Oberstufenprofilen

Die Tab. 2 zeigt die Mittelwerte der unterschiedlichen Leistungsmaße, aufgebrochen nach Oberstufenprofil und Erhebungszeitpunkt (im Falle der drei Leistungstests). Zur besseren Interpretation der Leistungsunterschiede sei darauf hingewiesen, dass sich bei einer 500/100-Metrik in früheren Arbeiten am Ende der Sekundarstufe I für das Fach Englisch gezeigt hat, dass der Kompetenzzuwachs eines Schuljahres bei 30 bis 40 Punkten liegt (Köller et al. 2010; Köller und Baumert 2018).

Erkennbar steigen die Mittelwerte in allen drei getesteten Kompetenzdimensionen von T1 nach T2 an. Die Zuwächse liegen zwischen 11 (Lesen in den sonstigen Profilen) und 41 (Hören im sprachlichen Profil) Punkten. Analog zu den Befunden bei Leucht et al. (2016) schneiden Schülerinnen und Schüler im sprachlichen Profil besser ab als in den übrigen Profilen. Diese Unterschiede nehmen noch etwas über die Zeit zu. Der Befund, wonach die Schülerinnen und Schüler des naturwissenschaftlichen Profils eine intellektuell positiv ausgelesene Gruppe bilden (höhere kognitive Grundfähigkeiten), ist bereits bei Köller (2016) für eine andere Stichprobe dokumentiert. Schließlich zeigt sich bei den Noten, dass diese im sprachlichen Profil am besten ausfallen, vor allem im Fach Englisch.

Ausführlicher soll im Folgenden auf die inferenzstatistischen Analysen in den Tab. 3–5 eingegangen werden. Dort wird auf der Basis regressionsanalytischer Befunde für die drei Teilkompetenzen und die beiden Messzeitpunkte getrennt berichtet, in welchem Ausmaß sich die Profile unterscheiden, ob die Differenzen nach Einführung von Kovariaten (kognitive Grundfähigkeiten und Geschlecht) stabil bleiben und ob sich Profilunterschiede auch bei der Veränderung (Modell 4 zu T2 in den drei Tabellen) nachweisen lassen. Die Regressionskoeffizienten der Profile lassen sich als Mittelwertsdifferenzen zum sprachlichen Profil interpretieren (Modell 1) bzw. als Mittelwertsdifferenzen nach Kontrolle der anderen Prädiktoren (Modelle 2 bis 4). Unterschiede zwischen den drei aufgeführten Profilen (Differenzen der Regressionsgewichte) wurden zusätzlich durch Wald-Tests (χ^2-Test mit $df = 1$) auf Signifikanz ($p < 0{,}05$) geprüft.

4.2.1 Profilunterschiede in der Lesekompetenz

Zu T1 zeigen sich zunächst die erwarteten Mittelwertsunterschiede, d. h. Schülerinnen und Schüler des sprachlichen Profils schnitten besser ab als die der übrigen Profile, wobei diese Differenz im Falle des naturwissenschaftlichen Profils nicht signifikant ist. Schülerinnen und Schüler des naturwissenschaftlichen Profils schnitten zudem signifikant besser ab als die der sonstigen Profile (Wald-Test: $\chi^2 = 3{,}904$, $df = 1$, $p < 0{,}05$). Werden die kognitiven Grundfähigkeiten als Kovariate eingeführt (Modell 2 zu T1), so steigt vor allem das negative Gewicht des naturwissenschaft-

Tab. 2 Testleistungen und Zeugnisnoten nach Profil und Messzeitpunkt (Mittelwerte/Standardabweichungen)

	Profil							
	Naturwissenschaftlich		Sprachlich		Gesellschaftswissenschaftlich		Sonstige	
	T1	T2	T1	T2	T1	T2	T1	T2
Lesen	501/104	514/100	519/87	538/92	492/79	509/94	471/81	482/96
Hören	497/106	518/114	520/103	561/100	494/92	517/106	480/84	495/103
Schreiben	495/107	509/112	530/88	553/92	490/95	509/105	471/96	485/110
Kognitive Grundfähigkeiten	529/108	–	478/93	–	488/93	–	484/88	–
Deutschnote	8,13/2,57	–	9,49/2,73	–	8,35/2,62	–	8,06/2,60	–
Englischnote	8,36/2,59	–	10,25/2,64	–	8,77/2,64	–	8,62/2,54	–
Mathematiknote	9,04/2,97	–	9,09/3,10	–	7,65/2,83	–	7,98/3,00	–

T1 Beginn der 11. Jahrgangsstufe, *T2* Ende der 11. Jahrgangsstufe; Mittelwert und Standardabweichung für die Leistungstests zu T1: $M = 500$, $SD = 100$; Eta-Quadrate (zwischen den Profilen): Lesen – T1 = 0,022; T2 = 0,024; Hören – T1 = 0,014; T2 = 0,035; Schreiben – T1 = 0,031; T2 = 0,038; Kognitive Grundfähigkeiten – T1 = 0,047; Deutschnote – T1 = 0,042; Englischnote – T1 = 0,072; Mathematiknote – T1 = 0,048

lichen Profils an und wird auch signifikant. Bei gleicher kognitiver Grundfähigkeit liegen also alle anderen Profile signifikant unter den Leistungen des sprachlichen Profils. Unter Berücksichtigung der Befunde in Tab. 2 ist diese Veränderung gegenüber Modell 1 plausibel, weist sie doch die im Mittel niedrigeren Werte des sprachlichen Profils in den kognitiven Grundfähigkeiten aus. Wird für diesen Malus korrigiert, steigen die Leistungsunterschiede im Leseverstehen. Die kognitiven Grundfähigkeiten selbst haben einen deutlichen Effekt und führen zu einem erheblichen Anstieg der aufgeklärten Varianz (R^2). Die Unterschiede zwischen dem naturwissenschaftlichen, dem gesellschaftswissenschaftlichen und den sonstigen Profilen sind in Modell 2 nicht signifikant.

In Modell 3 wird zusätzlich das Geschlecht als Kovariate eingeführt. Es zeigt sich kein Unterschied zwischen Schülerinnen und Schülern nach Kontrolle der übrigen Variablen. Ansonsten bleibt das Ergebnismuster gegenüber Modell 2 stabil.

Für den zweiten Erhebungszeitpunkt lässt sich festhalten, dass die Profildifferenzen zugunsten des sprachlichen Profils eher zugenommen haben (Modelle 1 bis 3) und jetzt auch der Effekt des Geschlechts signifikant wird (Modell 3), d. h. Schülerinnen schneiden besser ab als Schüler. Bei den Profilen lässt sich weiterhin festhalten, dass die sonstigen Profile signifikant schlechter als das gesellschaftswissenschaftliche Profil abschneiden (Modell 3 zu T2, Wald-Test: $\chi^2 = 3{,}919$, $df = 1$, $p < 0{,}05$).

Im letzten Schritt (Modell 4) wurde die Veränderung der Leseleistungen (T2 Lesen nach Kontrolle von T1 Lesen) vorhergesagt. Hier ist erkennbar, dass bei gleichen Ausgangswerten die Zuwächse im sprachlichen Profil höher waren als in den anderen Profilen, allerdings wurde nur die Differenz zu den sonstigen Profilen signifikant. Weiterhin ist erkennbar, dass Schülerinnen und Schüler mit steigenden kognitiven Grundfähigkeiten mehr dazulernten und Schülerinnen gegenüber Schülern höhere Kompetenzgewinne erreichten.

4.2.2 Profilunterschiede im Hörverstehen

Die regressionsanalytischen Befunde zum Hörverstehen (Tab. 4) weisen ein sehr ähnliches Muster auf wie die im Lesen. Allerdings werden die Abstände zwischen dem sprachlichen Profil und den übrigen Profilen zwischen T1 und T2 noch größer (Modell 1 zu T1 und T2), was letztendlich abbildet, dass die Zuwächse im sprachlichen Profil deutlich höher ausfallen als in den übrigen Profilen (Modell 4). Zu betonen ist hier, dass die Unterschiede in den Zuwächsen (bei Konstanthaltung der Ausgangsleistung) bei einer drittel Standardabweichung liegen. In nur neun Monaten zwischen beiden Erhebungszeitpunkten vergrößerten Schülerinnen und Schüler im sprachlichen Profil ihren Vorsprung um einen Betrag, der typischerweise dem Kompetenzzuwachs eines Schuljahres entspricht (s. oben; vgl. hierzu auch Köller und Baumert 2018).

Hinsichtlich der Geschlechtsdifferenzen zeigen sich zwar auch wieder zu T2 und in der Veränderung leichte Vorteile der Schülerinnen, diese erreichen aber keine Signifikanz. Schließlich ergeben sich wieder substanzielle Effekte der kognitiven Grundfähigkeiten in allen Vorhersagemodellen (Modelle 2 bis 4).

Tab. 3 Vorhersage von Lesekompetenzen im Fach Englisch; Befunde (unstandardisierte Regressionsgewichte und in Klammern Standardfehler) aus Regressionsanalysen (gemittelt über 15 Datensätze)

	Abhängige Variablen						
	T1 Lesen			T2 Lesen			
	Model 1	Model 2	Modell 3	Modell 1	Modell 2	Modell 3	Modell 4
Prädiktoren							
Profil (Referenz: sprachlich)							
Nawi	–18,0 (10,3)	–37,0** (10,4)	–34,8** (10,4)	24,0* (9,5)	–39,4** (9,6)	–31,6** (10,2)	–15,5 (9,1)
Wipo	–27,6** (9,0)	–31,6** (8,3)	–29,4** (8,9)	–28,9* (11,6)	–32,2** (10,8)	–24,3* (11,0)	–10,7 (10,5)
Sonst	–45,3** (12,2)	–49,1** (11,9)	–48,3** (11,8)	–51,7** (15,5)	–54,8** (14,6)	–52,2** (14,3)	–29,8* (13,2)
Kogn. Grundf	–	0,371** (0,044)	0,371** (0,044)	–	0,310** (0,039)	0,304** (0,039)	0,132** (0,040)
Geschlecht (Referenz: männlich)	–	–	6,0 (5,8)	–	–	21,3** (8,0)	18,5* (8,1)
T1 Lesen	–	–	–	–	–	–	0,463** (0,049)
R^2	0,022	0,178	0,179	0,024	0,116	0,127	0,284

T1 Beginn der 11. Jahrgangsstufe, *T2* Ende der 11. Jahrgangsstufe, *Nawi* naturwissenschaftliches Profil, *Wipo* gesellschaftswissenschaftliches Profil, *Sonst.* ästhetisches und sportwissenschaftliches Profil, *Kogn. Grundf.* Kognitive Grundfähigkeiten; * $p < 0{,}05$; ** $p < 0{,}01$, R^2 aufgeklärter Varianzanteil

Tab. 4 Vorhersage von Hörverstehenskompetenzen im Fach Englisch; Befunde (unstandardisierte Regressionsgewichte und in Klammern Standardfehler) aus Regressionsanalysen (gemittelt über 15 Datensätze)

	Abhängige Variablen						
	T1 Hören			T2 Hören			
	Model 1	Model 2	Modell 3	Modell 1	Modell 2	Modell 3	Modell 4
Prädiktoren							
Profil (Referenz: sprachlich)							
Nawi	–22,5 (11,8)	–44,3** (10,9)	–47,1** (11,2)	–43,3** (13,0)	–61,1** (12,4)	–56,8** (13,3)	–33,1* (13,7)
Wipo	–25,7* (10,4)	–30,3** (9,5)	–33,1** (9,7)	–44,2** (12,5)	–48,0** (11,9)	–43,7** (12,7)	–27,0* (13,2)
Sonst	–35,7* (14,4)	–40,1** (12,7)	–41,0** (12,7)	–59,6** (17,9)	–63,2** (16,8)	–61,8** (16,6)	–41,1* (16,1)
Kogn. Grundf	–	0,427** (0,036)	0,426** (0,036)	–	0,347 (0,046)	0,348** (0,045)	0,133* (0,055)
Geschlecht (Referenz: männlich)	–	–	–7,5 (6,9)	–	–	11,5 (8,9)	15,4 (8,3)
T1 Hören	–	–	–	–	–	–	0,503** (0,061)
R^2	0,014	0,187	0,189	0,035	0,131	0,134	0,306

T1 Beginn der 11. Jahrgangsstufe, *T2* Ende der 11. Jahrgangsstufe, *Nawi* naturwissenschaftliches Profil, *Wipo* gesellschaftswissenschaftliches Profil, *Sonst.* ästhetisches und sportwissenschaftliches Profil, *Kogn. Grundf.* Kognitive Grundfähigkeiten; * $p < 0{,}05$; ** $p < 0{,}01$, R^2 aufgeklärter Varianzanteil

Tab. 5 Vorhersage von Schreibkompetenzen im Fach Englisch; Befunde (unstandardisierte Regressionsgewichte und in Klammern Standardfehler) aus Regressionsanalysen (gemittelt über 15 Datensätze)

	Abhängige Variablen						
	T1 Schreiben			T2 Schreiben			
	Model 1	Model 2	Modell 3	Modell 1	Modell 2	Modell 3	Modell 4
Prädiktoren							
Profil (Referenz: sprachlich)							
Nawi	–35,0** (11,0)	–50,4** (10,2)	–50,7** (11,1)	–44,6** (12,0)	–61,8** (11,7)	–63,0** (12,5)	–24,8** (9,4)
Wipo	–40,0** (10,2)	–43,3** (9,5)	–43,6** (10,2)	–45,1** (11,6)	–48,8** (11,0)	–50,0** (11,7)	–17,1* (8,6)
Sonst	–52,7** (16,8)	–55,9** (15,1)	–56,0 (15,2)	–64,1** (18,6)	–67,6** (17,3)	–68,0** (17,4)	–35,8* (11,4)
Kogn. Grundf	–	0,302** (0,041)	0,301** (0,041)	–	0,336** (0,046)	0,336** (0,046)	0,109** (0,032)
Geschlecht (Referenz: männlich)	–	–	–0,82 (7,4)	–	–	–3,3 (8,3)	–2,7 (6,2)
T1 Schreiben	–	–	–	–	–	–	0,754** (0,037)
R^2	0,031	0,117	0,117	0,038	0,132	0,132	0,566

T1 Beginn der 11. Jahrgangsstufe, *T2* Ende der 11. Jahrgangsstufe, *Nawi* naturwissenschaftliches Profil, *Wipo* gesellschaftswissenschaftliches Profil, *Sonst.* ästhetisches und sportwissenschaftliches Profil, *Kogn. Grundf.* Kognitive Grundfähigkeiten; * $p < 0{,}05$; ** $p < 0{,}01$, R^2 aufgeklärter Varianzanteil

Tab. 6 Prozentuale Anteile der Schülerinnen und Schüler nach GER-Niveau und Profil in der Oberstufe

	A2		B1		B2		C1		C2	
	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
Gesamt	21,4	17,1	27,4	22,8	40,9	46,9	10,0	12,8	0,2	0,5
Nawi	25,8	19,7	25,1	20,0	36,6	47,5	12,2	12,5	0,3	0,3
Sprach	9,2	8,2	30,3	18,5	48,7	53,8	11,3	18,5	0,5	1,0
Wipo	22,7	19,2	30,0	26,5	40,4	42,7	6,9	11,5	0,0	0,0
Sonst	30,5	22,0	23,2	31,7	40,2	42,7	6,1	3,7	0,0	0,0

T1 Beginn der 11. Jahrgangsstufe, *T2* Ende der 11. Jahrgangsstufe, *Nawi* naturwissenschaftliches Profil, *Wipo* gesellschaftswissenschaftliches Profil, *Sonst.* ästhetisches und sportwissenschaftliches Profil

4.2.3 Profilunterschiede in den Schreibkompetenzen

Im Schreiben sind die Differenzen zwischen den Profilen noch einmal zugunsten des sprachlichen Profils größer und nehmen ebenfalls über die Zeit zu. Im Gegensatz zum Lesen ergaben sich keinerlei Geschlechterdifferenzen. Bei der Leistungsveränderung (Modell 4) zeigt sich wiederum ein starker Effekt zugunsten des sprachlichen Profils, erneut setzten sich diese Schülerinnen und Schüler um ein Lernjahr von den übrigen Schülerinnen und Schülern ab. Für die kognitiven Grundfähigkeiten ergab sich auch beim Schreiben, dass sie einen robusten Prädiktor der Fremdsprachenkompetenzen darstellen. Schließlich belegen das Regressionsgewicht der T1-Leistungen und der starke Anstieg im R^2 die hohe Stabilität der interindividuellen Differenzen in den Schreibkompetenzen.

Zusammenfassend zeigt sich mit Blick auf die Profilunterschiede, dass in allen drei Kompetenzbereichen Schülerinnen und Schüler des sprachlichen Profils höhere Werte erreichten als die der anderen Profile. Die Vorteile zugunsten des sprachlichen Profils sind dabei im Hörverstehen und Schreiben ausgeprägter als beim Lesen. Dies gilt auch für die Veränderung der Leistungen von T1 nach T2.

4.2.4 Schreibkompetenzen auf den Stufen des GER

Die abschließenden deskriptiven Analysen dienten der Verortung der festgestellten Schreibkompetenzen auf den Stufen A2 bis C2 des GER. Wie oben berichtet, wurde beim Standardsetting auf die Grenzsetzung zwischen A1 und A2 verzichtet, da die hier untersuchten Schülerinnen und Schüler nicht auf dem niedrigen A1-Niveau erwartet wurden. Mit der Verortung auf den GER-Stufen war es möglich, die Anteile der Schülerinnen und Schüler zu schätzen, die bereits in der 11. Jahrgangsstufe des achtjährigen Gymnasiums die normativen Ziele der Bildungsstandards für die Abiturprüfung (GER-Niveau B2) erreichen. Der Logik der Regelstandards folgend ist die Erwartung, dass die Hälfte der Schülerinnen und Schüler die gesetzten Ziele am Ende der Oberstufe erreicht. Die Tab. 6 zeichnet diesbezüglich ein sehr günstiges Bild. Bereits zu Beginn der 11. Jahrgangsstufe (T1) erreichen 51 % der Schülerinnen und Schüler mindestens das GER-Niveau B2 und damit die Vorgaben für das Erreichen der Allgemeinen Hochschulreife (KMK 2012). Am Ende des Schuljahres waren es dann sogar rund 60 %. Entsprechend den Befunden in Tab. 5 sind die Anteile der Schülerinnen und Schüler, die mindestens B2 erreichten, im sprachli-

chen Profil noch einmal deutlich höher (73,3 % zu T2) als in den übrigen Profilen, hier vor allem im Vergleich zu den ästhetischen/sportwissenschaftlichen (45,2 % zu T2; sonstige in Tab. 6). Die Tab. 6 zeigt aber auch, dass erhebliche Anteile der Schülerinnen und Schüler in der 11. Jahrgangsstufe nicht über die Stufe A2 hinauskommen. Zu T1 sind dies zwischen 9 (sprachliches Profil) und 30,5 % (ästhetisches/sportwissenschaftliches Profil), zu T2 liegen die Anteile zwischen 8 und 22 %.

5 Diskussion

Die vorliegende Untersuchung schließt an existierende empirische Untersuchungen zum Fremdsprachenlernen in der gymnasialen Oberstufe an (u. a. Köller und Trautwein 2004; Jonkmann et al. 2007, 2010; Leucht et al. 2016) und erweitert sie um die Analyse sprachproduktiver Kompetenzen im Fach Englisch. Das hier verwendete Paradigma zur Erfassung von Schreibkompetenzen wurde von ETS übernommen und nutzte zwei Sorten von Texten (*Independent* und *Integrated Tasks*), die sowohl von menschlichen Beurteilerinnen/Beurteilern (*Human Raters*) als auch von Computern (automatische Kodierung) ausgewertet wurden. Daneben wurden etablierte Tests zur Erfassung rezeptiver fremdsprachlicher Kompetenzen computerisiert erhoben. Die psychometrischen Analysen konnten die zufriedenstellende Qualität der verwendeten Instrumente belegen, dies galt vor allem für die verwendeten Schreibaufgaben. Inhaltlich wurde zum einen der Frage nach Unterschieden in den Leistungen zwischen unterschiedlichen Profilen in der gymnasialen Oberstufe nachgegangen. Zum anderen wurde für Schreibkompetenzen festgestellt, dass erhebliche Anteile der Schülerinnen und Schüler im ersten Jahr der Qualifikationsphase in der gymnasialen Oberstufe bereits normativ vorgegebene Ziele (GER-Stufe B2 oder höher) erreichen. Für die Kovariaten zeigte sich, dass die kognitiven Grundfähigkeiten ein robuster Prädiktor für die Leistungen und deren Veränderung waren, beim Geschlecht waren die Ergebnisse uneinheitlich. Im Folgenden gehen wir auf die Befunde zu den unterschiedlichen Fragestellungen ein und diskutieren sie. Dabei berücksichtigen wir Limitationen der vorliegenden Untersuchung und enden mit knappen Schlussfolgerungen.

5.1 Erfassung von Schreibkompetenzen

Während die Erfassung rezeptiver Kompetenzen im Fach Englisch in großen Schulleistungsuntersuchungen gut etabliert ist (Leucht et al. 2016; Stanat et al. 2016), werden sprachproduktive Leistungen in diesen Studien nicht berücksichtigt. Das hier verwendete *Multi-Level-Paradigma* zum Schreiben hat deutlich gemacht, dass es möglich ist, reliabel, valide und letztendlich auch ökonomisch (innerhalb von 60 min mit zwei Aufgaben) Schreibkompetenzen in der gymnasialen Oberstufe zu erfassen. Die Aufgaben ließen sich einfach am Computer bearbeiten und die Bewertungen (zwei *Human Raters* plus automatische Kodierung) ermöglichten die Zuweisung der Schülerinnen und Schüler zu unterschiedlichen Stufen des GER. Weiterhin erfassten sie Kompetenzen (Synthese von Informationen, argumentatives Schreiben), die typischerweise im Englischunterricht der gymnasialen Oberstufe the-

matisiert werden. Die deutlichen Korrelationen mit den Zeugnisnoten, welche höher ausfielen als die entsprechenden Koeffizienten für den Hör- und Lesetest, untermauern den engen Bezug zu den Anforderungen der Oberstufe. Kritisch mag man hier anführen, dass zwei Aufgaben zu wenig seien, um in der gymnasialen Oberstufe Schreibkompetenzen in ihrer curricularen Breite zu erfassen. Wir stimmen dieser Auffassung teilweise zu, sehen aber auch, dass zum einen die Ergebnisse zur Validität sehr ermutigend ausfielen und möchten zum anderen darauf hinweisen, dass auch beim Einsatz des TOEFL für die Zulassung an amerikanischen Universitäten Bewerberinnen und Bewerber nur zwei Schreibaufgaben im Rahmen der Gesamttestung bearbeiten müssen.

Aus unserer Sicht hat das hier verwendete Paradigma erhebliche Implikationen für zukünftige Schulleistungsuntersuchungen in Deutschland, wie sie beispielsweise vom IQB durchgeführt werden und die sich durch ihren *Low-Stakes-Charakter* auszeichnen. Für die von uns genutzten Texte liegen durch das Maschinenlernen Routinen für die automatische Kodierung der Texte vor. Würde man also in vergleichbaren Stichproben die Aufgaben erneut einsetzen, um Populationsparameter für die Leistungsverteilung und die Verteilung einer Zielpopulation auf Kompetenzstufen zu schätzen, so ließen sich sämtliche, durch Schülerinnen und Schüler produzierten Texte womöglich unter Verzicht auf menschliche Beurteiler auswerten. Damit wäre in der Tat die ökonomische Erfassung sprachproduktiver Fremdsprachenkompetenzen möglich. Wir sehen diese Chance natürlich nur auf aggregierter Ebene im Falle von *Low-Stakes-Assessments*, die keine Aussagen über individuelle Leistungsstände machen wollen. Assessments, die beispielsweise der Zugangsberechtigung zu Universitäten dienen, werden auch weiterhin auf *Human Ratings* zurückgreifen müssen. Ebenso ist es aus didaktischer Sicht unerlässlich, dass Textqualitäten wie „Adressatenorientierung“ oder „Argumentation“ in der Oberstufe ein hohes Gewicht behalten. Lehrpersonen müssen für die Arbeit im Klassenzimmer mit einer prozessorientierten Schreibdidaktik vertraut sein, welche auf motivierenden und anspruchsvollen Schreibaufgaben, Analysen von Mustertexten und gehaltvollen Rückmeldungen aufbaut. Im Falle von *Low-Stakes-Assessments* großer Populationen von Lernenden ist aber zu hoffen, dass mit dieser Arbeit Impulse für das automatische Kodieren von Schreibleistungen ausgehen.

5.2 Profilunterschiede

Analog zu früheren Untersuchungen (Leucht et al. 2015) ergaben sich substanzielle Unterschiede zwischen den Profilen in den Leistungen und Leistungszuwächsen, die alle zugunsten des sprachlichen Profils ausfielen. Die Befunde waren auch nach Kontrolle der Kovariaten Geschlecht und kognitive Grundfähigkeiten stabil, ja nahmen im Falle der Kontrolle kognitiver Grundfähigkeiten sogar noch zu. Ursache hierfür ist, dass die Schülerinnen und Schüler des sprachlichen Profils hinsichtlich ihrer kognitiven Grundfähigkeiten keineswegs positiv selegiert sind – anders als die des naturwissenschaftlichen Profils – und dementsprechend bei Konstanthaltung der kognitiven Grundfähigkeiten die Differenzen zugunsten des sprachlichen Profils zunahmen. Insgesamt zeigten sich beim Hören und beim Schreiben größere Differenzen, und zwar sowohl in den Mittelwerten zu T1 und T2 als auch in der Veränderung

der Leistungen von T1 nach T2. Bereits Leucht et al. (2015) haben Gründe für die Zunahme der Differenzen zwischen den Profilen (trotz gleicher Stundenzahl in der Oberstufe) diskutiert. Auch für unsere Befunde kann in diesem Sinne angeführt werden, dass die Vorteile im sprachlichen Profil sowohl schulisch als auch außerschulisch verursacht sein können; schulisch durch eine höhere Unterrichtsqualität, Transfereffekte der anderen belegten Fremdsprachen und eine günstigere Klassenkomposition; außerschulisch durch mehr fremdsprachliche Freizeitaktivitäten der Schülerinnen und Schüler des sprachlichen Profils.

5.3 Schreibkompetenzen und GER

Mit den Bildungsstandards für die fortgeführte Fremdsprache für die Allgemeine Hochschulreife (KMK 2012) liegen klare normative Erwartungen hinsichtlich des Erreichbaren vor. Während frühere Auswertungen für rezeptive Kompetenzen am Ende der Sekundarstufe II schon darauf hinwiesen, dass die normativen Zielvorgaben (B2/C1) von erheblichen Anteilen der Abiturientinnen und Abiturienten erreicht werden, fehlten entsprechende Befunde für die Schreibkompetenzen. Die vorliegende Untersuchung schließt diese Lücke teilweise. Wir konnten zeigen, dass bereits am Ende der 11. Jahrgangsstufe, also ein Jahr vor dem Abitur, rund 60 % der Schülerinnen und Schüler die Vorgaben erreichen, im sprachlichen Profil sogar fast drei Viertel aller getesteten Schülerinnen und Schüler. Da bei der Stichprobenziehung eine Repräsentativität für allgemeinbildende G8-Gymnasien in Schleswig-Holstein angestrebt wurde, haben wir keinen Zweifel, dass sich die Ergebnisse generalisieren lassen. Akzeptiert man zudem, dass in den Ländervergleichen des IQB für die Sekundarstufe I (zuletzt Stanat et al. 2016), Gymnasiastinnen und Gymnasiasten aus Schleswig-Holstein durchschnittlich (Lesen) bis leicht überdurchschnittlich (Hören) abschneiden, so lässt sich schlussfolgern, dass unsere Befunde auch relativ gut die Situation der Schreibkompetenzen national abbilden sollten. Insgesamt wird so ein optimistisches Bild der Erträge des gymnasialen Englischunterrichts gezeichnet. Großen Teilen der Schülerschaft gelingt es, argumentative Texte auf entsprechendem Niveau (B2/C1) zu schreiben. Der sehr stark kommunikativ und hier vor allem sprachproduktiv ausgerichtete Englischunterricht der gymnasialen Oberstufe, vermutlich gekoppelt mit vielen außerschulischen Fremdsprachenaktivitäten der Schülerinnen und Schüler, scheint hier die erhofften Wirkungen zu haben. Einschränkend muss hier konstatiert werden, dass in der vorliegenden Untersuchung keine Variablen erhoben wurden, um die vermuten Mediatoreffekte für die Vorteile im sprachlichen Profil zu testen.

5.4 Effekte der Kovariaten

Da sich die Oberstufenprofile in ihrer Komposition unterschieden, wurden basierend auf früheren Arbeiten (u. a. Leucht et al. 2016) die kognitiven Grundfähigkeiten und das Geschlecht als Kovariaten berücksichtigt. Für die kognitiven Grundfähigkeiten zeigte sich zum einen, dass sich bei ihrer Kontrolle die Leistungsvorteile im sprachlichen Profil verstärkten. Zum anderen erwiesen sich die kognitiven Grundfähigkeiten als robuster Prädiktor für die Englischleistungen und deren Veränderung. Die hohe

Bedeutung der kognitiven Grundfähigkeiten für Schulleistungen in allen Fächern ist ein äußerst robuster Befund, der erst kürzlich wieder in der Metaanalyse von Roth et al. (2015) zusammengefasst wurde. Schülerinnen und Schüler mit hohen kognitiven Grundfähigkeiten können Informationen schneller verarbeiten und kommen häufiger zu richtigen Schlussfolgerungen. Bemerkenswert in unserer Untersuchung war allerdings, dass die Effekte der kognitiven Grundfähigkeiten auch signifikant blieben, wenn das Vorwissen kontrolliert wurde (Modell 4 in den Tab. 3, 4 und 5). Zwar hatte das Vorwissen einen deutlich größeren Effekt, die Befunde zeigen aber, dass kognitive Grundfähigkeiten nicht nur einen leistungs- sondern einen lernförderlichen Effekt in der gymnasialen Oberstufe haben.

Die Veränderung der Profilunterschiede bei Berücksichtigung der kognitiven Grundfähigkeiten wurde oben bereits diskutiert. Dadurch, dass die Schülerinnen und Schüler des naturwissenschaftlichen Profils höhere kognitive Grundfähigkeiten aufwiesen als die der übrigen Profile und es einen positiven Effekt der kognitiven Grundfähigkeiten auf Leistungen gab, musste sichtbar werden, dass bei Kontrolle der kognitiven Grundfähigkeiten die Vorteile des sprachlichen gegenüber dem naturwissenschaftlichen Profil zunehmen.

Hinsichtlich der Geschlechterdifferenzen ergab sich kein einheitliches Bild, obwohl in der Literatur wiederholt Vorteile zugunsten der Mädchen berichtet werden (z. B. Winkelmann und Groeneveld 2010). Lediglich beim Leseverstehen zeigten sich zu T2 und in der Veränderung von T1 zu T2 signifikante Vorteile der Schülerinnen, die weder beim Zuhören noch beim Schreiben nachweisbar waren. Auch beim letzten Ländervergleich des IQB ergab sich an Gymnasien, allerdings am Ende der 9. Jahrgangsstufe, lediglich ein signifikanter Effekt zugunsten der Mädchen im Lesen, nicht aber im Hören (vgl. Böhme et al. 2016). Ursachen dafür werden dort nicht diskutiert. Denkbar wäre es, dass die allgemein stärkeren Leseaktivitäten von Mädchen außerhalb der Schule ihnen einen Vorteil im Bereich Leseverstehen bescheren. Solche rezeptiven Kompetenzen lassen sich jedoch nicht ohne weiteres auf produktive Bereiche übertragen. Man kann annehmen, dass besonders die Entwicklung der Fähigkeit zum argumentativen Schreiben in der Fremdsprache ein systematisches schulisches Training voraussetzt und nicht „nebenbei“ erworben wird (Keller 2013). Die fehlenden Geschlechterunterschiede beim englischen Schreiben wären dann ein positives Indiz dafür, dass es dem schulischen Fremdsprachenunterricht gelingt, beide Geschlechter anzusprechen und ausreichend zu fördern.

6 Schlussfolgerungen und Ausblick

Die Erfassung sprachproduktiver Leistungen in großen Stichproben stellte bislang eine erhebliche, nicht zuletzt ökonomische Herausforderung für die Schul- und Unterrichtsforschung dar. Mit dem hier vorgestellten Paradigma wurden Wege für die gymnasiale Oberstufe aufgezeigt, wie dieser Herausforderung zufriedenstellend begegnet werden kann. Es war möglich, mit jeweils zwei Schreibaufgaben die Leistungen der Schülerinnen und Schüler objektiv, reliabel und valide zu erfassen. Unsere Untersuchung sollte daher Anlass geben, die Forschungen zu den sprachproduktiven Kompetenzen der Schülerinnen und Schüler in deutschen Schulen zu verstärken.

Die automatisierte Kodierung dieser Leistungen mit intelligenten computerisierten Systemen wird fortschreiten und es scheint uns ein zentrales Desiderat zu sein, hier national entsprechende Systeme wie den E-Rater® und andere Systeme dieser Art zu entwickeln. Die Testentwicklung im Bereich der Fremdsprachenforschung sollte die Möglichkeiten in Zukunft viel stärker nutzen, um bei der Testung von Schülerinnen und Schüler ein breiteres Kompetenzspektrum erfassen zu können und damit auch mehr Akzeptanz in der Praxis zu schaffen. Nationale Vergleiche, die sich auf rezeptive Fremdsprachenkompetenzen beschränken, werden weder dem Fach Englisch gerecht noch können sie Impulse in Hinblick auf Sprachproduktion geben.

Vor allem für die Unterrichtsentwicklung ergeben sich bedeutende Zukunftsperspektiven. Je „intelligenter" solche Systeme wie E-Rater® werden, desto besser wird es möglich sein, eine Diagnose vorhandener Schreibkompetenzen zu geben, die Schülerinnen und Schüler aber auch direkt im Schreibprozess mit Informationen auszustatten (im Sinne formativen Assessments), wie sie ihre Kompetenzen steigern können. Die aktuell vorherrschende Praxis, wonach erst das fertige Schreibprodukt Gegenstand der Rückmeldung bzw. Bewertung ist, könnte so ergänzt werden durch eine individuelle Begleitung bzw. Unterstützung des Schreibprozesses.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz (http://creativecommons.org/licenses/by/4.0/deed.de) veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Anhang

Tab. 7 Überblick über die vier verwendeten Aufgaben

Aufgabe	Name	Inhalt
Independent	„Teachers"	A teacher's ability to relate well with students is more important than excellent knowledge of the subject being taught
	„Advertising"	Television advertising directed toward young children (aged two to five) should not be allowed
Integrated	„Voting Machines"	A text and lecture on the use of different voting systems used in the United States. Traditional systems were contrasted with computerized voting systems and the values of each discussed contrastively
	„The Chevalier"	A text and lecture on the memoirs of the Chevalier de Seingalt (1725–1798, more widely known as Giacomo Casanova). Participants were asked to discuss conflicting statements relating to the reliability of the Chevalier's memoirs, who liked „to make his life seem more exciting and glamorous than it really was"

Literatur

Beck, B., & Klieme, E. (Hrsg.). (2007). *Sprachliche Kompetenzen – Konzepte und Messung – DESI-Studie*. Weinheim: Beltz.

Böhme, K., Sebald, S., Weirich, S., & Stanat, P. (2016). Geschlechtsbezogene Disparitäten. In P. Stanat, K. Böhme, S. Schipolowski & N. Haag (Hrsg.), *IQB Bildungstrend 2015. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich* (S. 377–407). Münster: Waxmann.

Burstein, J., Kukich, K., Wolff, C., Lu, C., & Chodorow, M. (1998). *Enriching automated scoring using discourse marking* (Proceedings of the Workshop on Discourse Relations & Discourse Marking, Annual Meeting of the Association of Computational Linguistics, Montreal, Canada, August, 1998).

Cizek, G.J. (Hrsg.). (2012). *Setting performance standards: Foundations, methods, and innovations* (2. Aufl.). New York: Routledge.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002.

Europarat (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Berlin: Langenscheidt.

Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R.J., & Köller, O. (in press). Unlocking the TOEFL iBT rubrics for European writing assessment: Establishing a validity framework for cut scores in evidence-based standard setting. *Assessing Writing*.

Fleckenstein, J. (2017). *The common European framework of reference for languages: comparability and validity issues in assessment practice* (Dissertationsschrift). Kiel: Christian-Albrechts-Universität zu Kiel.

Flower, L., & Hayes, J.R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387. https://doi.org/10.2307/356600.

Grabe, W., & Kaplan, R.B. (1996). *Theory and practice of writing: an applied linguistic perspective*. New York: Longman.

Hambleton, R.H., Jaeger, R.M., Plake, B.S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, *24*(4), 355–366. https://doi.org/10.1177/01466210022031804.

Harsch, C., Lehmann, R., Neumann, A., & Schröder, K. (2007). Schreibfähigkeit. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie* (Bd. 1, S. 52–72). Weinheim: Beltz.

Harsch, C., Pant, H.A., & Köller, O. (Hrsg.). (2010). *Calibrating standards-based assessment tasks for English as a foreign language – standard-setting procedures in Germany*. Münster: Waxmann.

Harsch, C. (2006). *Der Gemeinsame europäische Referenzrahmen: Leistung und Grenzen. Die Bedeutung des Referenzrahmens im Kontext der Beurteilung von Sprachvermögen am Beispiel des semikreativen Schreibens im DESI-Projekt* (Inaugural-Dissertation). Augsburg: Universität Augsburg.

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, *9*, 122–159. https://doi.org/10.1016/j.asw.2004.06.001.

Heller, K.A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision*. Göttingen: Beltz.

Jonkmann, K., Köller, O., & Trautwein, U. (2007). Englischleistungen am Ende der Sekundarstufe II. In U. Trautwein, O. Köller, R. Lehmann & O. Lüdtke (Hrsg.), *Schulleistungen von Abiturienten. Regionale, schulformbezogene und soziale Disparitäten* (S. 113–142). Münster: Waxmann.

Jonkmann, K., Trautwein, U., Nagy, G., & Köller, O. (2010). Fremdsprachenkenntnisse in Englisch vor und nach der Neuordnung der gymnasialen Oberstufe in Baden-Württemberg. In U. Trautwein, M. Neumann, G. Nagy, O. Lüdtke & K. Maaz (Hrsg.), *Schulleistungen von Abiturienten. Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand* (S. 181–213). Wiesbaden: VS.

Keller, S. (2013). *Integrative Schreibdidaktik Englisch für die Sekundarstufe. Theorie, Prozessgestaltung, Empirie*. Tübingen: Narr.

KMK (2004). *Beschlüsse der Kultusministerkonferenz. Bildungsstandards für die erste Fremdsprache (Englisch/Französisch) für den Mittleren Schulabschluss. Beschluss vom 04. 12. 2003*. München: Luchterhand.

KMK (2012). *Bildungsstandards für die fortgeführte Fremdsprache (Englisch/Französisch) für die Allgemeine Hochschulreife. Beschluss der Kultusministerkonferenz von 18. 10. 2012*. Berlin: KMK.

Köller, O. (2016). Soziale Herkunft und kognitive Grundfähigkeiten der Schülerinnen und Schüler. In M. Leucht, N. Kampa & O. Köller (Hrsg.), *Fachleistungen beim Abitur. Vergleich allgemeinbildender und beruflicher Gymnasien in Schleswig-Holstein* (S. 99–17). Münster: Waxmann.

Köller, O., & Trautwein, U. (2004). Englischleistungen von Schülerinnen und Schülern an allgemein bildenden und beruflichen Gymnasien. In O. Köller, R. Watermann, U. Trautwein & O. Lüdtke, (Hrsg.), *Wege zur Hochschulreife in Baden-Württemberg. TOSCA – Eine Untersuchung an allgemeinbildenden und beruflichen Gymnasien* (S. 285–326). Opladen: Leske + Budrich.

Köller, O., & Baumert, J. (2018). Schulische Leistungen und ihre Messung. In W. Schneider & U. Lindenberger (Hrsg.), *Entwicklungspsychologie* (8. Aufl., S. 663–680). Weinheim: Beltz.

Köller, O., Baumert, J., Cortina, K. S., Trautwein, U., & Watermann, R. (2004). Öffnung von Bildungswegen in der Sekundarstufe II und die Wahrung von Standards: Analysen am Beispiel der Englischleistungen von Oberstufenschülern an integrierten Gesamtschulen, beruflichen und allgemeinbildenden Gymnasien. *Zeitschrift für Pädagogik, 50*, 679–700.

Köller, O., Knigge, M., & Tesch, B. (Hrsg.). (2010). *Sprachliche Kompetenzen im Ländervergleich*. Münster: Waxmann.

Leucht, M., & Köller, O. (2016). Anlage und Durchführung der Studie. In M. Leucht, N. Kampa & O. Köller (Hrsg.), *Fachleistungen beim Abitur. Vergleich allgemeinbildender und beruflicher Gymnasien in Schleswig-Holstein* (S. 79–98). Münster: Waxmann.

Leucht, M., Fleckenstein, J., & Köller, O. (2016). Erreichen kriterialer Leistungsstandards in der ersten Fremdsprache Englisch. In M. Leucht, N. Kampa & O. Köller (Hrsg.), *Fachleistungen beim Abitur. Vergleich allgemeinbildender und beruflicher Gymnasien in Schleswig-Holstein* (S. 171–199). Münster: Waxmann.

Leucht, M., Retelsdorf, J., Pant, H. A., Möller, J., & Köller, O. (2015). Effekte der Gymnasialprofilzugehörigkeit auf Leistungsentwicklungen im Fach Englisch. *Zeitschrift für Pädagogische Psychologie*, *29*(2), 77–88. https://doi.org/10.1024/1010-0652/a000153.

McCutchen, D. (2000). Knowledge, processing, and working memory: Implications for a theory of writing. *Educational Psychologist*, *35*(1), 13–23. https://doi.org/10.1207/S15326985EP3501_3.

Muthén, B., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267–316. https://doi.org/10.2307/271070.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8. Aufl., S. 1998). Los Angeles: Muthén & Muthén.

van Ockenburg, L., van Weijen, D., & Rijlaarsdam, G. (2016). Learning to write synthesis texts: a review of intervention studies. *Journal of Writing Research*, *10*(3), 402–428. https://doi.org/10.17239/jowr-2019.10.03.01.

Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing, 15*(2), 68–85. https://doi.org/10.1016/j.asw.2010.05.004

Plakans, L. (2010). Independent vs. integrated writing tasks: a comparison of task representation. *TESOL Quarterly*, *44*(1), 185–194. https://doi.org/10.5054/tq.2010.215251.

Porsch, R., & Köller, O. (2010). Standardbasiertes Testen von Schreibkompetenzen im Fach Englisch. In W. Bos, E. Klieme & O. Köller (Hrsg.), *Schulische Lerngelegenheiten und Kompetenzentwicklung* (S. 85–103). Münster: Waxmann.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). Evaluation of the e-rater® scoring engine for the TOEFL independent and integrated prompts. *ETS Research Report Series*. https://doi.org/10.1002/j.2333-8504.2013.tb02335.x. RR-12-06.

Reiss, K., Sälzer, C., Schiepe-Tiska, A., Klieme, E. & Köller, O. (Hrsg.). (2016). *PISA 2015: Eine Studie in Kontinuität und Wandel.* Münster: Waxmann.

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

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys. Wiley series in probability and mathematical statistics: applied probability and statistics*. Chichester: John Wiley & Sons Ltd.

Rupp, A. A., & Porsch, R. (2010). Standard-setting item pool. In C. Harsch, H. A. Pant & O. Köller (Hrsg.), *Calibrating standards-based assessment tasks for English as a first foreign language. Standard-setting procedures in Germany* (Bd. 2, S. 37–54). Münster: Waxmann.

Rupp, A. A., Casabianca, J., Fleckenstein, J., Keller, S., & Köller, O. (2019). Automated essay scoring at scale: a case study in Switzerland and Germany. In *TOEFL Research Report TOEFL-RR-86 and ETS Research Report; No. RR-19-12*. Princeton: Educational Testing Service. https://doi.org/10.1002/ets2.12249.

Schoonen, R. (2019). Are reading and writing building on the same skills? The relationship between reading and writing in L1 and EFL. *Reading and Writing*, *32*, 511–535. https://doi.org/10.1007/s11145-018-9874-1.

Stanat, P., Böhme, K., Schipolowski, S., & Haag, N. (Hrsg.). (2016). *IQB Bildungstrend 2015. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich*. Münster: Waxmann.

Tannenbaum, R. J., & Wylie, C. W. (2005). *Mapping test scores onto the common European framework*. Princeton, NJ: ETS. Setting standards of language proficiency on the Test of English as a Foreign Language (TOEFL), the Test of Spoken English (TSE), the Test of Written English (TWE), and the Test of English for International Communication

Winkelmann, H., & Groeneveld, I. (2010). Geschlechterdisparitäten. In O. Köller, M. Knigge & B. Tesch (Hrsg.), *Sprachliche Kompetenzen im Ländervergleich* (S. 177–184). Münster: Waxmann.