Schreibprodukte bewerten

- Die Rolle der Expertise bei der Bewertung der Textproduktionskompetenz

INAUGURALDISSERTATION

Zur Erlangung des Doktorgrades
der Humanwissenschaftlichen Fakultät
der Universität zu Köln
nach der Promotionsordnung vom 18.12.2018

vorgelegt von

Ann-Kathrin Hennes
aus Bonn

Köln, im März 2020

Für meine Eltern

– die mir beigebracht haben, dass ich alles schaffen kann.

Für meine Geschwister

– die bei allem, was ich geschafft habe, stets an meiner Seite waren.

Erstbetreuer der Arbeit: Prof. Dr. Alfred Schabmann (Universität zu Köln)

Zweitbetreuer der Arbeit: Prof. Dr. Christian Rietz (Pädagogische Hochschule Heidelberg)

Tag der mündlichen Prüfung: 06.07.2020

Diese Dissertation wurde von der Humanwissenschaftlichen Fakultät der Universität zu Köln im Juli 2020 angenommen.

Vorwort

Die vorliegende Dissertation hat die Form einer monografischen Dissertation mit Teilpublikationen, das heißt, einige der vorliegenden Textteile sind entweder bereits publiziert oder zur Publikation eingereicht worden. Dies trifft insbesondere auf die in Kapitel 5 und Kapitel 7 präsentierten Studien zu. Studie 1 der Dissertation (Kapitel 5) liegt (in englischer Sprache und revidierter Version) zur Begutachtung bei der Fachzeitschrift (*Journal of Writing Research*). Studie 2 dieser Arbeit (Kapitel 6) wurde bereits im Jahr 2018 unter dem Titel „Schreibkompetenz diagnostizieren: Ein standardisiertes Testverfahren für die Klassenstufen 4-9 in der Entwicklung“ in der Fachzeitschrift *Empirische Sonderpädagogik* (Nr. 3, S. 294-310) publiziert. Mit der freundlichen Genehmigung des Verlages wurden Teile dieser Publikation in ähnlicher Form in dieser Dissertation übernommen.

An Studie 1 (Kapitel 5) waren neben der Autorin folgende Autor*innen beteiligt: Barbara Maria Schmidt, Igor Osipov, Christian Rietz und Alfred Schabmann. Der Arbeitsbeitrag der Autorin dieser Dissertation ist jedoch im Vergleich zu den übrigen Autor*innen als überproportional einzuordnen, denn für alle Kapitel gilt, dass die Autorin diese hauptverantwortlich verfasst hat. Bei der Durchführung der statistischen Analysen war die Autorin proportional beteiligt, Gleiches gilt für die Konzeption des Studiendesigns. Die Autorin hat die Studie eigenverantwortlich durchgeführt.

Studie 2 (Kapitel 7) beziehungsweise die hierzu verwendeten Daten entstanden im Rahmen eines größeren interdisziplinären Kooperationsprojektes zur Konzeption eines standardisierten Schreibtests. Hieran waren neben der Autorin folgende Personen beteiligt: Michael Becker-Mrotzek, Jörg Jost, Markus Linnemann, Christian Rietz, Barbara Maria Schmidt und Alfred Schabmann. Der Autorin kam im Rahmen des Projektes die Aufgabe der Projektkoordination und Projektentwicklung zu. Das in Kapitel 7.1 vorgestellte Konvergenzmodell wurde zum Großteil von der Autorin dieser Dissertation entwickelt. Den publizierten Artikel verfasste die Autorin eigenständig, sie ist somit in überproportionalem Maße für die Inhalte verantwortlich. An der Durchführung der statistischen Analysen im Rahmen der Publikation war die Autorin proportional beteiligt.

Hinweis:
In dieser Arbeit wird aus Formulierungsgründen sowie zur besseren Lesbarkeit immer dann, wenn von Konstrukten (z.B. Experten und Adressatenorientierung) die Rede ist, die generische (männliche) Form verwendet. Gemeint ist dann sowohl die männliche als auch die weibliche Form.

Danksagung

Danke an:

Alfred, meinen Doktorvater und guten Berater, der mir in den letzten sechs Jahren gezeigt hat, wie ich schaffen kann, was ich möchte.

Christian, meinem Zweitbetreuer, der als mein Wegbereiter in die Wissenschaft danach gefragt hat, womit ich mich beschäftigen möchte.

Barbara, die als stille Beraterin meine ersten wackeligen Schritte supervidiert und mir dann beigebracht hat auf eigenen Füßen zu stehen.

Jörg, **Markus** und **Michael**, für die Kooperation auf Augenhöhe und die Erweiterung meines Blickwinkels.

Anna, **Annika**, **Alina**, **Barbara**, **Lisa**, **Lukas**, **Carla**, **Verena** und **Nora**, meine langjährigen Hilfskräfte, die mir viel Arbeit abgenommen und diese immer zuverlässig erledigt haben. Ihr habt mir den Fortschritt leichter gemacht.

Igor, für die vielen Diskussionen über die Möglichkeiten und Grenzen der Statistik.

Nico, der als mein langjähriger Lieblingskollege und Schreibtischnachbar immer ein Auge auf mich hatte.

Doris und **Inga**, die immer an einer Seite waren und für den notwendigen Spaß im Arbeitsleben gesorgt haben.

Kira, die mit ihrer Anwesenheit (und einer Menge Süßkram) viele anstrengende Tage und Aufgaben erträglicher gemacht hat.

Meine Freundinnen, die nie viel nachgefragt haben und mir so ein Leben neben der Dissertation und der Wissenschaft ermöglicht haben.

Eric, der mit seiner unermüdlichen Geduld und liebevollen Art dafür gesorgt hat, dass ich beim Schreiben dieser Arbeit nicht in der Phase der Überarbeitung stecken bleibe, sondern irgendwann aus dem Schreibprozess aussteige.

Meine Familie, für den festen Glauben an mich und die Vorfreude auf mein „Buch“ – hier ist es.

Alle Kinder, die Texte für mich geschrieben haben; **alle Lehrkräfte**, die mir ihre Unterrichtsstunden zur Verfügung gestellt haben und **alle Personen**, die an meinen Studien teilgenommen haben.

… und alle, die ich hier vergessen habe.

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

„[…] Wenn ich vorher gewußt hätt, was das für Mühe ist, 'n Buch zu machen, dann hätte ich mich gar nicht erst rangemacht und werd's auch nicht wieder machen" (Huckleberry Finn, Marc Twain, 2009, S.473).

1. Einleitung

Das Schreiben eines Textes in Form eines Buches, eines Briefes, einer Instruktion oder einer Dissertation (um nur einige Beispiele zu nennen) ist schwer und mühsam, denn die Produktion eines guten Textes ist eine komplexe Aufgabe (Allen, Snow & McNamara, 2016; Bereiter & Scardamalia, 1987). Schreibende stehen hierbei vor der Herausforderung, Gedanken- bzw. gedachte Informationsnetzwerke so in Schriftsprache zu übersetzen, dass Lesende diese rekonstruieren und somit nachvollziehen können. Zur Bewältigung dieses komplexen Prozesses bedarf es neben basaler Schreibfähigkeiten einer Vielzahl an hierarchiehohen Kompetenzen, denn (gute) Schreiber*innen müssen bei der Textproduktion die Bedürfnisse potentieller Adressaten berücksichtigen, den Prozess der Generierung, Auswahl und Organisation von Ideen bewältigen, über ausreichendes Anwendungswissen zu Textmustern verfügen und – je nach Text – sogar ihre Kreativität nutzen (Hennes et al., 2018). All diese Fähigkeiten dienen als Instrumente dazu, einen kohärenten Text zu erstellen, der den Bedürfnissen und Erwartungen der Lesenden gerecht wird und somit als *funktionaler* Text bezeichnet werden kann (Harsch, Neumann, Lehmann & Schröder, 2007; Hennes et al., 2018).

Die Kompetenz hierzu, die Textproduktionskompetenz[1], ist eine der wichtigsten Schlüsselfertigkeiten zur Teilhabe an der Gesellschaft sowie für schulischen und beruflichen Erfolg. Schüler*innen sollten diese daher im Laufe der Pflichtschulzeit erwerben und hierbei durch guten Schreibunterricht (und bei Bedarf durch individuelle Förderung) unterstützt werden. Wichtigste Voraussetzung für einen guten Schreibunterricht und individuelle Fördermaßnahmen bildet eine objektive, reliable und valide Diagnostik.

Forschungsbefunde zur diagnostischen Kompetenz von Lehrkräften im Bereich der Textproduktionskompetenz (siehe Kapitel 4) zeigen jedoch, dass Lehrkräfte Probleme damit haben, objektive und zuverlässige Bewertungen der Textqualität, welche derzeit als Maß für die Textproduktionskompetenz herangezogen wird, vorzunehmen. Verschiedene Lehrkräfte bewerten ein und denselben Text unterschiedlich und damit weder objektiv noch reliabel und folglich

[1] Der oben beschriebene Satz an hierarchiehohen Fähigkeiten wird häufig als *Schreibkompetenz* bezeichnet. In dieser Arbeit wird jedoch der Begriff der *Textproduktionskompetenz* (siehe z.B. Baer, Fuchs, Reber-Wyss, Jurt & Nussbaum, 1995) verwendet, um das Konstrukt, auf dem der Schwerpunkt dieser Arbeit liegt, von sehr grundlegenden Schreibfähigkeiten, wie beispielsweise der Rechtschreibung, abzugrenzen (siehe Kapitel 2).

auch nicht valide. Eine zuverlässige diagnostische Grundlage für die Anpassung des Unterrichts an die Bedürfnisse der Schüler*innen und die Förderung schreibschwacher Schüler*innen steht Lehrkräften somit nicht zur Verfügung.

Mögliche Ursachen für die mangelnde Übereinstimmung zwischen Lehrkräften bei der Bewertung der Textproduktionskompetenz wurden bisher nur in geringem Umfang beforscht. Existierende Studien, welche als mögliche Ursache hauptsächlich die mangelnde diagnostische Expertise der Lehrkräfte im Bereich der Schreibkompetenz diskutieren (siehe z.B. Birkel, 2003), weisen aber immer auch auf ein grundsätzliches Problem hin: Es fehlen einheitliche und eindeutige Textbewertungskriterien zur Anwendung im schulischen Kontext; und auch im wissenschaftlichen Kontext existieren keine manifesten Kriterien zur validen Beurteilung der Textqualität.

Ein weiteres Problem, welches sich an dieser Stelle aus förderdiagnostischer Perspektive ergibt, ist die Tatsache, dass die Textqualität nur bedingt Aussagen zur Textproduktionskompetenz zulässt. Ist ein Text beispielsweise nicht funktionstüchtig, kann dies unterschiedliche Gründe haben. Zum Beispiel kann die/der Schreiber*in beim Verfassen des Textes nicht in der Lage gewesen sein, wichtige Informationen als solche zu identifizieren, oder Schwierigkeiten dabei gehabt haben, die Aufgabestellung zu verstehen und eine entsprechende Zielsetzung vorzunehmen. In beiden Fällen wird das Geschriebene nicht nachvollziehbar (funktionsuntüchtig) und damit von schlechter Qualität sein. Worin genau die Schwierigkeiten bei der Textproduktion lagen, lässt sich anhand der (globalen) Textqualität zumeist nicht genauer bestimmen. Zur systematischen Erfassung der einzelnen Komponenten der Textproduktionskompetenz bedarf es entsprechend differenzierter diagnostischer Verfahren. Ein förderdiagnostisch ausgerichtetes Verfahren, welches psychometrischen Gütekriterien gerecht wird, existiert im deutschen Sprachraum bisher nicht. Didaktisch orientierte Bewertungsraster, die verschiedene Aspekte der Textualität unabhängig voneinander fokussieren, stellen jedoch erste Ansatzpunkte für eine differenzierte Erfassung der Textproduktionskompetenz dar.

Ziel der vorliegenden Dissertation ist es, einen Beitrag zur Verbesserung der Diagnostik im Bereich der Textproduktionskompetenz zu leisten. Hierzu soll zunächst die Frage geklärt werden, welche Rolle die Expertise bei der Textbewertung spielt und wie die diagnostische Expertise von Lehrkräften in diesem Bereich (verglichen mit Experten und Novizen des Feldes) einzuordnen ist. Im Rahmen der hierzu durchgeführten Studie (Studie 1 der Dissertation; Kapitel 5) soll geklärt werden, ob es Lehrkräften „lediglich“ an Expertise zur Textbewertung mangelt (mit welcher sie entlang von Aus- und Weiterbildung ausgestattet werden könnten) oder ob

das Fehlen eindeutiger Bewertungskriterien die valide Bewertung von Texten bzw. die valide Diagnostik von Textproduktionskompetenzen im Allgemeinen und nicht nur für Lehrkräfte erschwert beziehungsweise einschränkt.

In einem zweiten Schritt soll dann der aktuelle Stand der Schreibforschung diskutiert werden (Kapitel 6). Aufbauend hierauf wird in Studie 2 (Kapitel 7) dieser Dissertation ein alternativer, psychometrischen Kriterien entsprechender Ansatz zur Beurteilung der Textproduktionskompetenz und ihrer einzelnen Dimensionen vorgestellt. Das diesem Ansatz zugrundeliegende Modell zum Konstrukt der Textproduktionskompetenz und erste statistische Eigenschaften des entwickelten Instrumentes werden präsentiert. Darüber hinaus werden besondere Herausforderungen, die bei der Konstruktion eines solchen Verfahrens zu berücksichtigen sind, thematisiert. Zunächst soll jedoch das Konstrukt, welchem sich diese Arbeit widmet, genauer definiert werden (Kapitel 2).

2. Textproduktionskompetenz

Eine einheitliche Definition dazu, was unter Schreibkompetenzen (hier Textproduktionskompetenzen) konkret zu verstehen ist, existiert in der wissenschaftlichen Fachliteratur nicht.

Eine der umfangreichsten Konstruktdefinitionen findet sich in den Bildungsstandards der Kultusministerkonferenz (KMK). Hier wird die Schreibkompetenz als Fertigkeit dazu definiert, „Texte dem Zweck entsprechend und adressatengerecht gestalten, sinnvoll aufbauen und strukturieren [zu können]“ (KMK, 2003, S. 11). Spezifikationen dieser Definition werden vorgenommen und beziehen sich auf Aspekte der sprachlichen Gestaltung eines Textes (Strukturierung, Verständlichkeit, sprachliche Variabilität, Einsatz sprachlicher Mittel wie beispielsweise Vergleiche) sowie inhaltliche Komponenten (beispielsweise das sinnvolle Gewichten von Argumenten und das Ziehen von logischen Schlüssen). Außerdem werden Aspekte sprachlicher Richtigkeit (Orthografie und Grammatik) in die Definition eingeschlossen. Insgesamt sollen Schüler*innen zur Erlangung eines mittleren Schulabschlusses hinsichtlich dreier Anforderungsbereiche ausgebildet werden. Hierzu zählen laut KMK (2003): 1) die Transkription, 2) die Rechtschreibung und Zeichensetzung und 3) der Prozess der Textproduktion (Planung, Übersetzung und Überarbeitung). Ganz ähnlich wird die Schreibkompetenz auch im internationalen Bildungskontext definiert. Die US-amerikanische Institution NAEP (National Assessment of Educational Progress) versteht unter der Schreibkompetenz die Fähigkeit dazu, einen Text zu produzieren, welcher orthografisch und grammatikalisch richtig, in einem angemessenen Stil

und zusammenhängend geschrieben ist und in welchem alle relevanten/notwendigen Informationen gegeben werden (NAEP, 2011). Beide Definitionen rücken die Produktion eines funktionalen Textes, der bestimmte Kriterien erfüllt, in den Mittelpunkt. Sie sind damit am entstehenden Text, d.h. dem Produkt des Schreibprozesses, orientiert.

Alternative Definitionen beziehen sich auf kognitive Komponenten der Textproduktionskompetenz und betrachten die Schreibkompetenz als eine Art Kompetenzbündel aus verschiedenen Teilkompetenzen, die zur Produktion von funktionstüchtigen Texten benötigt werden. Diese Definitionen gehen nicht vom Textprodukt aus, sondern setzen an den notwendigen Fähigkeiten und Kompetenzen zur Produktion derselben an (z.B. Bachmann & Becker-Mrotzek, 2017; Becker-Mrotzek, 2014; Becker-Mrotzek & Böttcher, 2012; Knopp, Jost, Nachtwei, Becker-Mrotzek & Grabowski, 2012). Ein Beispiel hierfür ist die Definition nach Becker-Mrotzek und Schindler (2007; 2008). Hier wird die Schreibkompetenz in vier Wissenstypen (deklaratives Wissen, Problemlösewissen, prozedurales Wissen und metakognitives Wissen) unterteilt, welche entlang definierter Anforderungsbereiche der Textproduktion (Medium, Orthografie, Lexik, Syntax, Textmuster und Leserorientierung) operationalisiert werden. Entlang dieser Definition braucht die/der Schreibende zur Produktion eines funktionstüchtigen Textes in allen genannten Anforderungsbereichen deklaratives Wissen (d.h. Faktenwissen, z.B. darüber, welche Textmuster es gibt und wie diese Muster aufgebaut sind), Problemlösewissen (d.h. methodische Kompetenzen, z.B. die Fähigkeit, ein bekanntes Textmuster mit Inhalt zu füllen und entsprechend der Zielsetzung des Textes adaptieren zu können), prozedurales Wissen (d.h. die Kompetenz, über Prozeduren und Routinen verfügen zu können und diese entsprechend einzusetzen, z.B. Wissen um und Anwendung von standardisierten Formulierungen und Strukturierungen, welche zum Textmuster passen) sowie die Kompetenz zur Selbstregulation und Selbstüberwachung (metakognitive und strategische Kompetenzen). Die (im Idealfall) vorhandenen Kompetenzen in den Bereichen der vorgestellten vier Wissenstypen[2] gilt es dann, entsprechend den Anforderungen des Textproduktionsvorhabens, zielführend miteinander zu kombinieren. Hierbei ist zu berücksichtigen, dass verschiedene Schreibanlässe bzw. Schreibaufgaben unterschiedliche Anforderungen an die schreibende Person stellen können (z.B. durch Unterschiede in der Komplexität des jeweils geforderten Textmusters). Die Schreibkompetenz lässt sich der vorliegenden Definition nach somit „als das Produkt aus Anforderungsniveau der Schreibaufgabe und der Summe des anforderungsbezogenen Wissens definieren“ (Becker-Mrotzek &

[2] Die Autoren verwenden bei der Darstellung ihres Kompetenzmodells ausschließlich den Begriff des „Wissens“, beziehen sich in ihren Darstellungen im eigentlichen Sinne jedoch zum Großteil auf Kompetenzen (siehe oben).

Schindler, 2008, S. 99). Neben dem zu schreibenden Text liegt in dieser Definition der Fokus auf den hierzu (notwendigen) Kompetenzen der Schreiber*innen.

Trotz der benannten Unterschiede in der Ausrichtung der in diesem Kapitel dargestellten Definitionen lassen sich auch Überschneidungen zwischen diesen im Hinblick auf das Konstrukt der Textproduktionskompetenz finden. Mit Blick auf die erforderlichen Kompetenzen von Schreibenden, auf deren Erfassung die Bewertung der Textqualität ausgerichtet ist, lassen sich verschiedene Aspekte benennen, welche zur Definition des Konstruktes der Textproduktionskompetenz herangezogen werden können. Hierzu zählen:

- Die Fähigkeit zur *Adressatenorientierung.* (siehe auch: Becker-Mrotzek & Schindler, 2007; Becker-Mrotzek & Böttcher, 2012; Bereiter, 1980; Ehlich, 1983; Hayes & Flower, 1980; Hayes, 2012; Becker-Mrotzek, Grabowski, Jost, Knopp & Linnemann, 2014; Scardamalia & Bereiter, 1987; Schmitt & Knopp, 2017).
- Das *Textmusterwissen* und die Fähigkeit, Textmuster (samt zugehöriger Routinen und Formulierungen) anzuwenden und, wenn nötig, adaptieren zu können (siehe auch: Bereiter, 1980; Heinemann, 2000; Hayes & Flower 1980; Hayes, 2012; Knopp, Jost, Linnemann & Becker-Mrotzek, 2014; Sandig, 1997).
- Die Fähigkeit, notwendige Inhalte/Informationen bestimmen zu können und diese sinnvoll gewichten sowie logisch darstellen zu können, d.h. ein gelungenes *Informationsmanagement* („Informationslogistik"; Hennes et al., 2018, S. 299) betreiben zu können (siehe auch: Bereiter & Scardamalia, 1987; Hayes & Flower, 1980; Hayes, 2012).
- Die Fähigkeit, einen Text verständlich und gut strukturiert verfassen zu können und diesem einen inneren Zusammenhang zu geben, d.h. *Kohärenz* herstellen zu können (siehe auch: Averintseva-Klisch, 2013; Graesser, McNamara, Louwerse & Cai, 2004; MacArthur, Jennings & Philippakos, 2019).
- Die Fähigkeit, einen Text sprachlich ansprechend zu gestalten (z.B. durch sprachliche Variabilität und die Verwendung von Vergleichen), d.h. über *sprachliche Kreativität* zu verfügen (siehe auch: Dem-Jensen & Heine, 2013; Kellogg, 2008).
- Die Fähigkeit zur Bewältigung des Schreibprozesses, d.h. Planungs-, Transkriptions- und Überarbeitungsphasen sinnvoll aufeinander abzustimmen (siehe auch Hayes & Flower, 1980; Hayes, 2012).

- Die Fähigkeit, Überwachungs- und Regulationsprozesse während der Textproduktion erfolgreich zu steuern und dabei die Zielsetzung im Blick behalten (siehe auch: Hayes & Flower, 1980; Hayes, 2012; Alamargot, Caporossi, Chesnet & Ros, 2011)
- Die Fähigkeit, orthografisch und grammatikalisch richtig zu schreiben.

Die benannten Aspekte (oder Teile derselben) werden in der Literatur häufig als Teilkomponenten der Textproduktionskompetenz modelliert (Becker-Mrotzek et al., 2014), welche es im Kontext des Schreibprozesses sinnvoll aufeinander abzustimmen gilt. Diese Abstimmung stellt, wie eingangs erwähnt, eine komplexe Herausforderung dar. Die Textproduktion wird daher von Schreibenden häufig als mühsam empfunden. Dies gilt nicht nur für Schreibanfänger*innen, sondern auch für fortgeschrittene Schreiber*innen. Im folgenden Kapitel wird deutlich: Es ist wichtig, über (ein Mindestmaß an) Textproduktionskompetenz zu verfügen (Hennes et al., 2018), auch wenn ihr Erwerb viel Fleiß und Zeit erfordert (Kellogg, 2008).

3. Zur Relevanz der Textproduktionskompetenz und der Notwendigkeit guter Diagnostik

Die Textproduktionskompetenz ist ein wichtiger Prädiktor für schulischen (Crossley & McNamara, 2016; Feenstra, 2014; Graham & Perin, 2007; Koster, Tribushinina, De Jong & Van den Bergh, 2015; National Commission on Writing, 2003; Becker-Mrotzek, 2014) sowie beruflichen Erfolg (McNamara, Knoch & Fan, 2019). Im Schulkontext hilft das Schreiben, neue Informationen zu strukturieren und zu vernetzen (Graham & Hebert, 2011; Linnemann & Stephany, 2014), was wiederum eine wichtige Voraussetzung für das Lernen ist (Farnan & Fearn, 2008; Pohl & Steinhoff, 2010). Zudem erfordern Leistungsüberprüfungen häufig die Produktion eines Textes (MacArthur, Graham & Fitzgerald, 2016). Später dann, in der Arbeitswelt, erfordern viele Arbeitsplätze ein Mindestmaß an Textproduktionskompetenz (MacArthur et al., 2016). Dies gilt auch für kognitiv weniger anspruchsvolle Arbeitsplätze, die z.B. für Menschen mit Lernschwierigkeiten geeignet sind (Bach, Schmidt, Schabmann & van Kroll, 2016). Schüler*innen, die Textproduktionskompetenzen nicht erwerben und daher nur über mangelnde Fähigkeiten zur Textproduktion verfügen, sind in ihrer gesellschaftlichen Teilhabe beeinträchtigt.

Aufgrund ihrer Bedeutung sollte die Vermittlung von Textproduktionskompetenzen notwendiger Bestandteil des Schreibunterrichts und Basisziel möglichst vieler schulischer Bildungsgänge sein (National Commission on Writing, 2003; Hodges, Wright, Wind, Matthews,

Zimmer & McTigue, 2019; Hennes et al., 2018). Lehrpersonen sollten Textproduktionskompetenzen zielgerichtet vermitteln und ihren Unterricht entsprechend der individuellen Lernbedürfnisse der Schüler*innen gestalten können (Graham & Perin, 2007; Graham, McKeown, Kiuhara & Harris, 2012; Beck, Llosa, Black & Anderson, 2018).

Wichtige Voraussetzung dafür ist, dass Lehrkräfte die vorhandenen Kompetenzen ihrer Schüler*innen im Bereich der Textproduktion zutreffend beurteilen können und über differenziertes Wissen zu den individuellen Stärken und Schwächen ihrer Schüler*innen im Bereich der Textproduktion verfügen. Mit anderen Worten: Lehrkräfte müssen über ein entsprechendes Maß an Expertise im Bereich der Diagnostik von Textproduktionskompetenzen verfügen (Beck et. al., 2018; Hodges et al., 2019).

Ein wesentliches Element diagnostischer Expertise ist theoretisches Wissen zu dem zu beurteilenden Konstrukt (Hesse & Latzko, 2017). Als Experten müssen Lehrkräfte daher die verschiedenen Komponenten der Textproduktionskompetenz kennen, um diese anhand adäquater Kriterien bewerten zu können. Nur so können adäquate Schlussfolgerungen für den Unterricht oder gegebenenfalls notwendige Interventionen gezogen werden. Hat ein(e) Schüler*in beispielsweise bereits die Fähigkeit, die notwendigen Informationen beim Schreiben eines Textes auswählen zu können, kann diese aber nicht kohärent präsentieren, dann sollte der Unterricht darauf ausgerichtet sein, Möglichkeiten der Kohärenzstiftung zu vermitteln, anstatt die Auswahl relevanter Informationen zu fokussieren.

4. Aktuelles Vorgehen zur Diagnostik von Textproduktionskompetenzen im schulischen Kontext

Im schulischen Kontext wird die Textproduktionskompetenz von Schüler*innen aktuell zumeist anhand eines längeren Textes erhoben (Feenstra, 2014; Philipp, 2015) und in Form eines „Gesamturteils“ zur Textqualität beurteilt (Eckes, 2008, S. 160). Teilweise werden zusätzlich analytische Kriterien wie der Inhalt, die Organisation oder bestimmte sprachliche Merkmale (z.B. Wortschatz, Satzarten und Grammatik; vgl. Eckes, 2008; Weigel, 2002) zur Textbewertung herangezogen.

Ein solches Vorgehen erscheint naheliegend, da die Schreibkompetenz im Rahmen der Bildungsstandards über Merkmale gelungenerer Schreibprodukte definiert wird (siehe Kapitel 2). Empirische Studien zeigen jedoch, dass die auf diese Weise gewonnenen Urteile von geringer Reliabilität und damit auch von geringer Validität sind. So zeigten Birkel und Birkel (2002),

dass sich die Bewertungen verschiedener Lehrkräfte zu ein und demselben Text stark voneinander unterschieden; Bewertungen für den gleichen Text variierten über das gesamte Notenspektrum, d.h. zwischen den Noten 1 (sehr gut) und 6 (mangelhaft). Ähnliche Befunde berichten Cooksey, Freebody und Wyatt-Smith (2007). Sie fanden relativ hohe Inkonsistenzen in den Textbewertungen von Lehrkräften, wobei der Anteil der Varianz in den Bewertungen, welcher sich durch die Lehrkräfte und nicht durch die Textqualität erklären ließ, zwischen 30% und 70% lag. Inkonsistenzen ergaben sich vor allem dann, wenn Texte von durchschnittlicher Qualität bewertet wurden, während sehr gute und sehr schlechte Schreibprodukte konsistenter bewertet wurden (Cooksey et al., 2007).

Als Ursachen für die fehlenden Übereinstimmungen zwischen Lehrkräften und damit die unzureichende Zuverlässigkeit und Validität von Lehrkrafturteilen in Bezug auf die Textqualität werden im aktuellen Forschungskontext verschiedene Faktoren diskutiert:

1) Es werden Unterschiede in der Interpretation und Gewichtung von Textbewertungskriterien angeführt (Eckes, 2008), welche sich entweder darin zeigen, dass verschiedene Lehrkräfte die gleichen Kriterien unterschiedlich auslegen und/oder gewichten, oder aber darin, dass dieselbe Lehrkraft die gleichen Kriterien bei der Bewertung verschiedener Schüler*innentexte unterschiedlich auslegt und/oder gewichtet (Cooksey et al., 2007; Leckie & Baird, 2011). In beiden Fällen ergibt sich hieraus ein gravierendes Problem: Lehrkräfte nutzen unterschiedliche Maßstäbe zur Bestimmung der Textqualität und messen damit zwangsläufig unterschiedliche Konstrukte (Murphy & Yancey, 2008; Olinghouse, Santangelo & Wilson, 2012). Die Bewertungen verschiedener Lehrkräfte zum gleichen Text sind folglich nicht miteinander vergleichbar.
2) Als Ursache für die unzureichende Validität der Lehrkrafturteile wird die Tendenz der Lehrkräfte diskutiert, in erster Linie Kriterien zu verwenden, welche leicht anzuwenden sind und gleichzeitig eine Art „Augenscheinvalidität“ aufweisen. Empirische Befunde zeigen, dass Lehrkräfte häufig lediglich recht grundlegende Eigenschaften eines Textes bewerten (z.B. Rechtschreibung, Grammatik und Wortschatz, welche nicht genügen, um einen guten Text zu produzieren) und hierarchiehohen Kriterien, die in einem engeren Zusammenhang zur Textqualität und der Textproduktionskompetenz stehen (z.B. Textkohärenz), weniger Beachtung schenken (Birkel & Birkel, 2002; Murphy & Yancey, 2008; Rezaei & Lovorn, 2010; Smith, Cheville & Hillocks, 2006; Vögelin, Jansen, Keller & Möller, 2018; Vögelin, Jansen, Keller, Machts & Möller, 2019). Auf diese Weise vorgenommene Bewertungen sagen allerdings nur wenig über die Textproduktionskompetenz der Schüler*innen aus und sind somit nicht valide.

3) Als weiterer Faktor, der sich ungünstig auf die Zuverlässigkeit von Lehrkrafturteilen auswirkt, wird die Orientierung von Lehrkräften am Leistungsniveau der entsprechenden Klassen als Anker/Referenzrahmen für die Bewertung einzelner Schüler*innentexte diskutiert (Ingenkamp & Lissman, 2008; Cumming, Kantor & Powers, 2001; Cumming, Kantor & Powers, 2002) – ein Vorgehen, welches zu relativ guten (im Sinne von reliablen) Ergebnissen führt, solange Lehrkräfte die Qualität von Schüler*innentexten innerhalb des klasseninternen Bezugsrahmens (d.h. entlang der soziale Norm) bewerten (Cooksey et al., 2007), welches aber höchst unzureichend ist für die Beurteilung der „absoluten" Qualität eines Textes (d.h. entlang der kriterialen Norm und in Bezug zu den Texten aller Schüler*innen in einer bestimmten Population). Das Leistungsniveau einer Klasse ist somit entscheidend für die Bewertung der Textproduktionskompetenzen der Schüler*innen. Ein und die-/derselbe Schüler*in kann bei diesem Vorgehen sowohl als „schlechte/r" als auch als „gute/r" Schreiber*in bewertet werden – je nach Leistungsniveau der Bezugsklasse. Eine zuverlässige Identifikation von schreibschwachen Schüler*innen ist auf diese Weise nicht möglich; dies wiederum erschwert beziehungsweise verhindert die Durchführung notwendiger Interventionsmaßnahmen (Boone, Thys, van Avermaet & van Houtte, 2018; Marsh, 1987; Trautwein, Lüdtke, Marsh, Köller & Baumert, 2006). Ein (klassen-)unabhängiger und objektiver „Maßstab" in Form eindeutiger und zuverlässig anwendbarer Kriterien bildet hierfür eine zwingende Voraussetzung.

Eine genauere Betrachtung der aktuell existierenden Textbewertungskriterien zeigt jedoch, dass diese (z.B. Adressatenorientierung, gutes Informationsmanagement, oder sprachliche Kreativität) keinesfalls gut operationalisiert sind (Knoch, 2011; Lumley, 2002; Todd, Thienpermpool & Keyuravong, 2004). Vielmehr existiert keine klare Vorstellung dazu, was genau es bedeutet, *sich an den Bedürfnissen der Adressaten zu orientieren*, *ein gutes Informationsmanagement zu betreiben*, *Kenntnisse zu geeigneten Textmustern zu haben und diese einsetzen zu können*, *Kohärenz herzustellen* oder *sprachlich kreativ* zu sein (siehe Kapitel 2). Es mangelt an empirischen Befunden dazu, wie sich die einzelnen Komponenten der Textproduktionskompetenz operationalisieren lassen (siehe hierzu Kapitel 6.4). In Anbetracht dieser Situation erscheint es nachvollziehbar, dass Lehrkräfte die gleichen Kriterien unterschiedlich auslegen, bei der Bewertung auf Kriterien zurückgreifen, die sie mit gewisser Sicherheit anwenden können (z.B. Rechtschreibung), oder die Textproduktionskompetenz ihrer Schüler*innen anhand von Vergleichsprozessen bewerten.

Dennoch muss festgehalten werden, dass die berichteten Befunde darauf hinweisen, dass die auf diese Weise gewonnenen Einschätzungen der Lehrkräfte zur Textqualität nur wenig

zuverlässig zu sein scheinen. Bisher ist allerdings unklar, ob der Grund dafür in der herausfordernden Aufgabe selbst (der Bewertung der Textqualität) liegt, oder ob es Lehrkräften an Kompetenz fehlt (z.B. dem Wissen um adäquate Bewertungsmaßstäbe), um zuverlässige Bewertungen zur Textqualität vornehmen zu können – während andere Personengruppen (z.B. Lektor*innen und Autor*innen) über entsprechende Expertise bei der Textbewertung verfügen. Hierbei handelt es sich um eine wichtige Frage, denn wenn die Textbewertungen anderer Experten zuverlässiger sind als die Einschätzungen von Lehrkräften, dann wäre die Verbesserung deren diagnostischer Kompetenzen „lediglich" eine Frage der Lehrer*innenbildung – denn es gäbe Experten, die bereits über entsprechende Beurteilungskompetenzen verfügten, entlang welcher Lehrkräfte unterrichtet werden könnten. Wenn andererseits jedoch keine Unterschiede zwischen der Zuverlässigkeit der Bewertungen durch Lehrkräfte und der anderer Experten bestehen, würde sich ein grundlegenderes Problem bei der Bewertung der Textqualität (und damit auch für die Ausbildung von Lehrkräften im Hinblick auf die Textbewertung) ergeben. In diesem Fall würde sich dann die Frage stellen, ob überhaupt Experten für die Beurteilung der Qualität von (Schüler*innen-)Texten existieren und wenn ja, wer diese sind und nach welchen (möglicherweise impliziten) Kriterien sie die Textqualität beurteilen.

In der aktuellen Forschungslandschaft existieren bisher keine Daten[3], die einen Vergleich der diagnostischen Expertise von Lehrkräften mit der Expertise anderer potentieller „Expertengruppen" (siehe Kapitel 5.4 zur Diskussion dazu, welche Personengruppen als „Experten" betrachtet werden können) und damit eine Beantwortung der oben gestellten Fragen zulassen. Folglich ist bisher unklar, wie gravierend der Mangel an Zuverlässigkeit in den Textbewertungen von Lehrkräften im Vergleich zu anderen Experten (und Novizen) ist. Im Rahmen der in Kapitel 5 beschriebenen Studie wurden entsprechende Daten erhoben, die einen Vergleich zwischen Lehrkräften, potentiellen Experten und Novizen im Hinblick auf deren Expertise bei der Textbewertung zulassen.

5. Diagnostische Kompetenzen im Vergleich – Studie 1

An dieser Stelle setzt die im Folgenden dargestellte Studie (Studie 1 der Dissertation) an. Anhand des Experten-Novizen-Paradigmas (Voss, Fincher-Kiefer, Green & Post, 1986) werden die diagnostischen Kompetenzen von Lehrkräften bei der Textbewertung mit denen möglicher Experten, aber auch den Kompetenzen von Novizen verglichen.

[3] Nach bestem Wissen der Autorin

5.1 Das Experten-Novizen-Paradigma

Das Experten-Novizen-Paradigma stammt ursprünglich aus der Kreativitätsforschung (Amabile, 1982), wurde in der Vergangenheit aber auch im Kontext von Forschungsfragen zu Lehrkraftkompetenzen angewandt (z.B. Bromme, 2008). Das Paradigma bietet einen Rahmen für die Bewertung komplexer (kreativer) Produkte/Konstrukte, wenn hierzu keine klaren Kriterien vorliegen – so wie es entlang der bisherigen Argumentation für die Bewertung der Textqualität der Fall ist. Grundannahme des Paradigmas ist, dass Experten bei der Bewertung kreativer Produkte aufgrund ihrer Erfahrung und ihres Beitrags beziehungsweise Erfolges im entsprechenden Gebiet implizite Kenntnisse über entscheidende Faktoren und Kriterien guter kreativer Produkte haben (z.B. Kaufman, Baer, Cropley, Reiter-Palmon & Sinnett, 2013).

Verschiedene Studien, in denen das Experten-Novizen-Paradigma auf unterschiedliche kreative Bereiche (z.B. Filmproduktionen, Gemälde, Tortendekorationen, aber auch die Textproduktion) angewandt wurde, zeigen, dass sich Experten in ihren Bewertungen hauptsächlich in drei Punkten von Novizen unterscheiden.

Diese sind:

1) Die Strenge: Bei der Bewertung desselben Produktes geben Novizen tendenziell bessere Bewertungen als Experten ab (Kaufman, Baer, Cole & Sexton 2008; Kaufmann, Baer, Cole, 2009). Die Bewertungen von Novizen sind somit weniger streng (weniger kritisch) als die von Experten.

2) Die (Urteils-)Konsistenz: Experten, die ein und dasselbe Produkt beurteilen, sind sich über dessen Qualität weitgehend einig. Sie stimmen miteinander überein, d.h. ihre Bewertungen sind konsistent (Amabile, 1996; Kaufman et al., 2013). Im Gegenteil dazu sind die Bewertungen von Novizen recht inkonsistent (Kaufmann et al., 2008; Kaufman et al., 2013).

3) Die konsistente Anwendung und Gewichtung von relevanten Bewertungskriterien: Neben den beiden oben genannten Kriterien von Expertise kann als drittes Experten-Merkmal das Wissen um und die konsistente Anwendung von relevanten Bewertungskriterien genannt werden (Ben-Simon & Bennett, 2007). Denn die konsistente Verwendung und Gewichtung von Kriterien bei der Beurteilung der Qualität eines (kreativen) Produktes stellt eine mögliche Erklärung für die hohe Übereinstimmung zwischen den Bewertungen von Experten dar (Eckes, 2008; Jonsson & Svingby, 2007; Weigel, 1994).

5.2 Zielsetzung der Studie

Ziel der vorliegenden Studie ist es, die Expertise von Lehrkräften bei der Bewertung von (Schüler*innen-)Texten anhand der skizzierten Kriterien von Expertise zu untersuchen. Lehrkräfte werden hierzu mit Personengruppen, die im Sinne des Experten-Novizen-Paradigmas als Experten (beziehungsweise Novizen) betrachtet werden können, verglichen.

Die Gruppe der Experten bilden dabei zum einen Lektor*innen und Autor*innen (Expertengruppe Gruppe 1) und zum anderen Wissenschaftler*innen (Expertengruppe 2). Beiden Gruppen kann aufgrund ihrer Erfahrung mit Texten und deren Bewertung beziehungsweise deren Überarbeitung sowie ihrer Reputation im Feld Expertise zugeschrieben werden (für mehr Details zu den Versuchsteilnehmer*innen, deren Charakteristika und Auswahlkriterien siehe Kapitel 5.4).

5.3 Forschungsfragen

Entlang der in Kapitel 5.1 aufgeführten Kennzeichen von Expertise lauten die Forschungsfragen der vorliegenden Studie wie folgt:

1) Wie „streng“ sind die Textbewertungen von Lehrkräften im Vergleich zu den Bewertungen anderer (potentieller) Experten und Novizen?

Die Strenge einer Person wird anhand der von ihr abgegebenen Urteile zur Gesamtqualität der vorgegebenen Texte ermittelt (für statistische Details siehe Kapitel 5.4.4). Wie oben beschrieben, bewerten Experten kreative Produkte in der Regel strenger als Novizen, d.h. kreative Produkte werden von Experten im Mittel weniger gut bewertet als von Novizen.

Aufgrund der Tatsache, dass bisherige Forschungsbefunde nur geringe Übereinstimmungen (und damit eine große Varianz) zwischen den Urteilen verschiedener Lehrkräfte nachweisen konnten, wird erwartet, dass die Textbewertungen der Lehrkräfte im Mittel weniger streng als die der Experten, aber wahrscheinlich strenger als die der Novizen sind.

2) Wie konsistent sind die Textbewertungen der Lehrkräfte im Vergleich zu den Bewertungen anderer (potentieller) Experten und Novizen?

Gruppen, die (entlang des Experten-Novizen-Paradigmas) als Experten betrachtet werden können, weisen (entsprechend statistischer Kriterien) eine hohe Konsistenz innerhalb ihrer Bewertungen auf (für statistische Details siehe Kapitel 5.4.4). Lehrkräfte können demnach dann als Experten betrachtet werden, wenn sich eine hohe Konsistenz zwischen den Textbewertungen

verschiedener Lehrkräfte zeigt. Die Konsistenz der Lehrkrafturteile sollte sich dann nicht von der anderer Experten unterscheiden und höher sein als die von Novizen.

Aufgrund der in Kapitel 4 angeführten empirischen Befunde wird erwartet, dass die Konsistenz der Lehrkrafturteile geringer ist als die anderer Experten, aber höher ausfällt als bei Novizen.

3) Wie konsistent sind die Gewichtungen von Textbewertungskriterien bei Lehrkräften im Vergleich zu anderen (möglichen) Experten und Novizen?

Gruppen, die (entlang des Experten-Novizen-Paradigmas) als Experten betrachtet werden, weisen (entsprechend statistischer Kriterien) eine hohe Konsistenz innerhalb ihrer Gewichtungen der Textbewertungskriterien auf. Lehrkräfte können demnach dann als Experten betrachtet werden, wenn sich eine hohe Konsistenz innerhalb der Gewichtungen verschiedener Lehrkräfte zeigt. Die Konsistenz der Gewichtungen innerhalb der Gruppe der Lehrkräfte sollte sich nicht von der anderer Experten unterscheiden und höher sein als die bei Novizen. Auch hier wird aufgrund der aufgeführten Befunde erwartet, dass Lehrkräfte weniger konsistent in der Gewichtungen der Bewertungskriterien sind als andere Experten, aber wahrscheinlich konsistenter als Novizen.

4) Welches Gewicht geben Lehrkräfte relevanten Bewertungskriterien bei der Beurteilung der Textqualität im Vergleich zu anderen (möglichen) Experten und Novizen?

Gruppen, die (entlang des Experten-Novizen-Paradigmas) als Experten betrachtet werden können, sollten relevante Bewertungskriterien bei der Beurteilung der Textqualität berücksichtigen, d.h. diesen ein entsprechend hohes Gewicht zuschreiben; weniger relevanten Kriterien sollte ein geringeres Gewicht zugeschrieben werden. Lehrkräfte können dann als Experten der Textbewertung betrachtet werden, wenn die von ihnen vorgenommenen Gewichtungen der Bewertungskriterien ein entsprechendes Muster zeigen, d.h. relevanten Kriterien ein entsprechend höheres Gewicht zugeordnet wird als weniger relevanten.

Aufgrund der in Kapitel 4 aufgeführten Befunde kann angenommen werden, dass Lehrkräfte bei der Textbewertung neben relevanten Kriterien auch weniger relevante Kriterien berücksichtigen. Inwiefern sich diese Tatsache in den Gewichtungen der relevanten/weniger relevanten Kriterien widerspiegelt, gilt es zu untersuchen.

5.4 Methode

5.4.1 Stichprobe

Im Rahmen der Studie werden die Textbewertungen und Gewichtungen der Textbewertungskriterien von N=175 Experten, Lehrkräften und Novizen miteinander verglichen.

Entlang des Experten-Novizen-Paradigmas (z.B. Kaufmann et al., 2008; Amabile, 1982) kann einer Person dann der Experten-Status zugeschrieben werden, wenn sie (1) einen eigenen und allgemein akzeptierten Beitrag zum entsprechenden Gebiet geleistet hat und (2) über genügend Erfahrung/Wissen verfügt, auf das sie sich bei der Bewertung von (kreativen) Produkten beziehen kann. Auf Basis dieser Kriterien wurden zwei Gruppen von Experten identifiziert, deren Textbewertungen und Gewichtungen der Kriterien mit denen von Lehrkräften und Novizen verglichen wurden. Die Stichprobe setzt sich wie folgt zusammen:

Experten 1: Die erste Expertengruppe umfasst N=31 Lektor*innen und Autor*innen, die Kinder- und Jugendbücher lektorieren und/oder schreiben. Von den Personen dieser Gruppe waren 87,1% weiblich, das Durchschnittsalter lag bei 42,3 Jahren (*SD*=10,5). Als Lektor*innen arbeiteten 35,5% der Personen dieser Gruppe, 29,0% als Autor*innen und 35,5% sowohl als Lektor*innen als auch als Autor*innen. Ihre durchschnittliche Berufserfahrung betrug 10,8 Jahre (*SD*=6,8), im Durchschnitt fielen 56,5% (*SD*=34,7) ihrer Arbeitszeit in den Bereich der Kinder- und Jugendliteratur.

Experten 2: Die zweite Expertengruppe besteht aus Wissenschaftler*innen, die im Bereich der Schreibdidaktik tätig sind. Die Stichprobe umfasst 24 Personen (60,0% weiblich) mit einem Durchschnittsalter von 39,6 Jahren (*SD*=9,0). In dieser Stichprobe sind 32,0% Doktorand*innen, 44,0% Postdoktorand*innen und 24,0% Professor*innen. Die durchschnittliche Berufserfahrung in der Forschung betrug zum Erhebungszeitpunkt 11,0 Jahre (*SD*=7,7). Die befragten Personen gaben an, durchschnittlich 36,8% (*SD*=21,5) ihrer Arbeitszeit mit Texten, die für oder von Kindern und Jugendlichen geschrieben wurden, zu verbringen. Die Forschungsschwerpunkte aller Teilnehmer*innen bezogen sich auf die Beurteilung der Textproduktionskompetenz/die Bewertung von (Schüler*innen-)Texten.

Lehrkräfte: Die Vergleichsgruppe der Lehrkräfte umfasst N=51 berufstätige Grundschullehrer*innen (94,1% weiblich), die zum Zeitpunkt der Studie in Klasse 4 Deutsch unterrichteten oder innerhalb der letzten drei Jahre eine vierte Klasse in Deutsch unterrichtet hatten (dies war ein Auswahlkriterium für die Teilnahme an der Studie, da die zu bewertenden Texte

von Viertklässler*innen geschrieben wurden). Im Durchschnitt waren die teilnehmenden Lehrer*innen 12,4 Jahre als Lehrkraft tätig (*SD*=10,0), ihr Durchschnittsalter betrug 38,9 Jahre (*SD*=11,0).

Novizen: Die Gruppe der Novizen besteht aus 69 Universitätsstudierenden (91,3% weiblich) ohne Vorkenntnisse oder Erfahrung im Hinblick auf die Bewertung von (Schüler*innen-)Texten. Das Durchschnittsalter der Novizen lag zum Erhebungszeitpunkt bei 24,4 Jahren (*SD*= 3,3).

5.4.2 Auswahl und Rekrutierung der Versuchspersonen

Voraussetzung für die Teilnahme der Lektor*innen und Autor*innen an der vorliegenden Studie war ein klarer Arbeitsschwerpunkt im Bereich der Kinder- und Jugendliteratur. Ein Teil der teilnehmenden Lektor*innen/Autor*innen (N=10) wurde über den *deutschen Verband freier Lektoren* über die Studie informiert und zur Teilnahme aufgefordert (Verband freier Lektoren; https://www.vfll.de/). Die übrigen 21 Personen wurden anhand einer Internetrecherche ausgewählt und persönlich von der Autorin dieser Dissertation per E-Mail kontaktiert. Alle Lektor*innen/Autor*innen erhielten eine von ihnen selbst festgelegte Aufwandsentschädigung für die Teilnahme an der Befragung (die Höhe dieser variierte zwischen 0 und 130€).

Die Auswahl der Wissenschaftler*innen erfolgte über eine intensive Internetrecherche (Universitätshomepages) und eine Sichtung der relevanten nationalen Literatur zum Thema der Textbewertung/Schreibkompetenz. Potentielle Teilnehmer*innen wurden persönlich per E-Mail kontaktiert, über die Studie informiert und hinsichtlich ihrer Bereitschaft zur Teilnahme befragt.

Um Lehrkräfte für die Studie zu rekrutieren, wurden Schulleitungen von Grundschulen per E-Mail und Telefon über die Studie informiert. Diese wurden darum gebeten, Lehrkräfte ihrer Schulen zur Teilnahme zu ermutigen.

Novizen wurden im Rahmen von Lehrveranstaltungen an der Universität durch die Autorin dieser Arbeit rekrutiert.

Alle Teilnehmer*innen wurden über das Ziel der Studie informiert. Sie nahmen freiwillig teil und gaben ihr Einverständnis, dass die gewonnenen Daten für Forschungszwecke verwendet werden dürfen.

5.4.3 Instrumente

Die Teilnehmer*innen aller Gruppen wurden im Rahmen der Studie dazu aufgefordert, 20 Texte unterschiedlicher Qualität zu bewerten. Das Online-Befragungstool *soscisurvey* (www.soscisurvey.de) wurde für die Textbewertung, das Kriterien-Rating und den zugehörigen (demographischen) Fragebogen genutzt. Die Teilnehmer*innen wurden per E-Mail zur Teilnahme an der Studie eingeladen und erhielten einen Link mit zugehörigem Passwort, wodurch ihnen der Zugang zur Umfrage ermöglicht wurde. Die Befragungen wurden zwischen August 2016 und Juni 2018 durchgeführt.

5.4.3.1 Das Textkorpus

Das Studiendesign sieht vor, dass alle Teilnehmer*innen die Qualität der gleichen 20 Schüler*innentexte (N=20) bewerten. Zur Gewährleistung ausreichender Varianz zwischen diesen 20 Texten in Bezug auf die Textqualität wurde eine repräsentative Stichprobe aus einem größeren Textkorpus gezogen. Das für die Stichprobenziehung verwendete Textkorpus wurde auf folgende Weise erzeugt: Viertklässler (N=401) verschiedener Schulen in Nordrhein-Westfalen (Deutschland) wurden gebeten, einen kurzen narrativen Text auf Basis einer Bildergeschichte mit sechs aufeinanderfolgenden Bildern zu verfassen (siehe Abbildung 1). Alle Schüler*innen erhielten exakt den gleichen Schreibauftrag. Die (handgeschriebenen) Texte der Schüler*innen wurden dann in ein Textverarbeitungsprogramm eingegeben und auf Rechtschreibfehler korrigiert, um die Möglichkeit auszuschließen, dass Teilnehmer*innen die Texte anhand von Rechtschreibfehlern bewerten. Alle anderen Elemente des Textes (wie Überschriften und Absätze) wurden unverändert beibehalten.

Abbildung 1. Bildergeschichte der Schreibaufgabe (©2020, Orell Füssli Sicherheitsdruck AG, Globi Verlag, Imprint Orell Füssli Verlag, Zürich, Abbildung aus Papa Moll, Band 10, PM die Sportskanone)

Mit dem Ziel, den Arbeitsaufwand für den folgenden Arbeitsschritt zu reduzieren, wurde aus dem so aufbereiteten Textkorpus eine Zufallsstichprobe von 201 Texten gezogen. Diese 201 zufällig ausgewählten Texte wurden anschließend anonym anhand der Methode des *Comparative Judgment* (CJ; Pollitt, 2012) bewertet. Bei diesem Vorgehen handelt es sich um eine in der Schreibforschung etablierte Methodik (siehe van Daal, Lesterhuis, Coertjens, Donche & De Maeyer, 2016). Im Rahmen des CJ wurden Studierenden der Universität zu Köln (N=47 Lehramtsstudierende mit Bachelor-Abschluss; Durchschnittsalter 25,1 Jahre; 87,2% weiblich) jeweils zwei Texte präsentiert, welche sie direkt miteinander vergleichen sollten. Auf Basis des vorgenommenen Vergleiches sollten die Studierenden dann für jedes (Text-)Paar entscheiden, welcher der beiden Texte besser ist. Auf diese Weise wurde jeder der 201 Texte im Durchschnitt zwanzig Mal mit einem anderen Text verglichen. Anhand der aus den Vergleichen gewonnenen Daten wurde dann nach einem von Bradley und Terry (1952) vorgeschlagenen Verfahren für jeden Text ein Logit-Score berechnet. Dieser Score spiegelt für jeden Text die Wahrscheinlichkeit wider, den Vergleich mit einem anderen Text zu gewinnen (Politt, 2012) und kann somit als Indikator für die Textqualität verwendet werden (Whitehouse & Politt, 2012). Die Reliabilität der gebildeten Skala (*Scale-Separation-Reliability*; siehe Bramley, 2015) lag bei 0,81, d.h.

die für die 201 Texte berechneten Logit-Scores wiesen eine zufriedenstellende interne Konsistenz auf. Alle Vergleiche und Berechnungen wurden mittels des Softwarpakets *D-PAC* durchgeführt (Lesterhuis, Verhavert, Coertjens, Donche & De Maeyer, 2017).

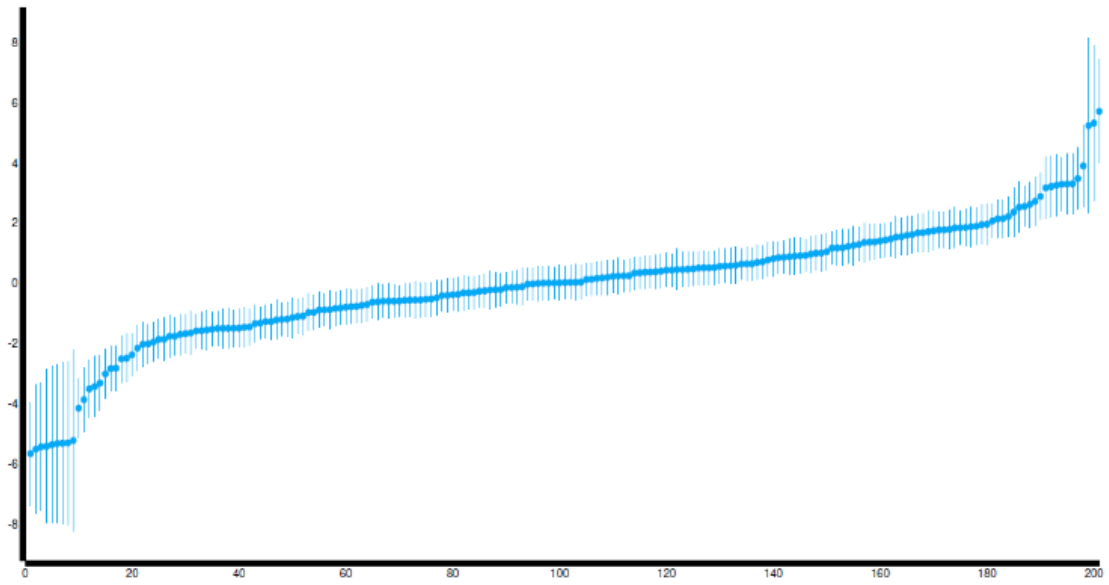


Abbildung 2: Verteilungskurve der Texte entlang der ermittelten Logit-Scores;
Anmerkung: Auf der horizontalen Achse sind die Logit-Scores abgebildet; auf der vertikalen Achse ist die Stichprobe, hier die Texte, abgebildet.

In einem letzten Schritt wurden die Texte dann anhand ihrer Logit-Scores in 20 Perzentile unterteilt. Aus jedem der 20 Perzentile wurde zufällig ein Text gezogen, um das Material auf die gewünschten 20 Texte (von unterschiedlicher Qualität) zu reduzieren. Die so ausgewählten 20 Schüler*innentexte wurden dann für das im Folgenden beschriebene Textbewertungsverfahren verwendet.

5.4.3.2 Befragung zur Textbewertung

Alle Teilnehmer*innen wurden gebeten, die Qualität der (gleichen) 20 Schüler*innentexte anhand eines Globalurteils auf einer sechsstufigen Skala von 1 (sehr schlecht) bis 6 (sehr gut) zu bewerten. Ein solches Vorgehen ist sowohl in der Forschung (z.B. Birkel & Birkel, 2002; Cooksey et al., 2007; Crossely & McNamara, 2010) als auch im Schulkontext üblich (Noten zwischen 1 „sehr gut“ und 6 „ungenügend“ können für die Gesamtqualität eines Textes vergeben werden).

Vor der Bewertung der Texte wurden die Teilnehmer*innen über das Alter und die Klassenstufe der Schreiber*innen informiert. Außerdem erhielten sie Informationen zur vorgegebenen Schreibaufgabe und zur vorgenommenen Rechtschreibkorrektur. Spezifische Angaben zur Identität der jeweiligen Schreiber*innen (z.B. Geschlecht oder nationale Herkunft) wurden nicht gemacht.

Im Zuge der Textbewertung sahen die Teilnehmer*innen aller Gruppen 20 Fragebogenseiten (siehe Abbildung 3), die alle gleich aufgebaut waren. Im oberen Teil der Seite stand die Frage: „Wie würden Sie den folgenden Text als Ganzes beurteilen?“ Darunter wurde einer der 20 Schüler*innentexte präsentiert. Unter dem jeweiligen Text war dann die sechsstufige Bewertungsskala abgebildet. Die Teilnehmer*innen konnten ihre Bewertungen abgeben, indem sie auf den entsprechenden Punkt der Bewertungsskala klickten.

Abbildung 3: Beispielseite des Fragebogens zur globalen Textbewertung

Die Schüler*innentexte wurden in zufälliger Reihenfolge präsentiert, um Reihenfolgeeffekte auszuschließen. Die Bewertenden konnten die Texte nicht direkt miteinander vergleichen. Das Vor- und Zurückblättern zwischen den Fragebogenseiten war nicht möglich. Diese Einstellung wurde vorgenommen, um zu verhindern, dass die Teilnehmer*innen die Texte auf Basis von Vergleichen bewerteten. Die Bewertenden wurden so dazu gezwungen, die „absolute“ Qualität jedes Textes zu bewerten.

5.3.3.3 Gewichtung einzelner Dimensionen der Textproduktionskompetenz

Nach Abschluss aller 20 globalen Textbewertungen wurden die Teilnehmer*innen gebeten, vier Dimensionen von Textproduktionskompetenz hinsichtlich ihrer Bedeutung für die vorgenommenen globalen Textbewertungen zu gewichten.

Die zu bewertenden Dimensionen (Literatur siehe Kapitel 2) waren:

1) Adressatenorientierung,

2) Informationsmanagement,

3) Einsatz geeigneter Textmuster(aspekte) und

4) die Herstellung von Kohärenz.

Jede Dimension wurde anhand einer Reihe von Kriterien operationalisiert (siehe Anhang A). Zum Beispiel wurde die Dimension *Einsatz geeigneter Textmuster(aspekte)* (hier die Erzählung) anhand der folgenden Kriterien operationalisiert: a) der Text hat eine Einleitung, b) der Text hat einen Schluss, c) der Text ist im richtigen Tempus geschrieben, d) im zeitlichen Verlauf des Textes wird Spannung aufgebaut, e) durch den Einsatz der wörtlichen Rede im Text wird ein szenischer Charakter erzeugt, f) der Text enthält emotionale Momente und g) die Lesenden werden durch den Aufbau des Textes beim Lesen unterstützt. Alle zu gewichtenden Kriterien wurden entlang der wissenschaftlichen Literatur (z.B. Averintseva-Klisch, 2013; Becker-Mrotzek, 2014; Graesser et al., 2004; Hayes & Flower, 1980) zu den Komponenten der Schreibkompetenz (hier Textproduktionskompetenz genannt) ausgewählt. Die Formulierungen basieren auf bestehenden Bewertungsskalen (vgl. Eckes, 2008; Hachmeister, 2019; Nussbaumer, 1991; Thonke, Groß Ophoff, Hosenfeld & Isaac, 2008), aber auch auf einer im Vorfeld der Studie durchgeführten Pilotbefragung von Lehrkräften dazu, welche Kriterien sie bei der Beurteilung der Textqualität berücksichtigen. Aus den Ergebnissen der Befragung wurden alternative Formulierungen der Kriterien abgeleitet. Dieser Schritt erschien notwendig, um Alternativen zu wissenschaftlichen Formulierungen zu finden und damit Kriterien zu formulieren, die für alle Gruppen gleichermaßen verständlich waren.

Darüber hinaus wurden fünf Kriterien in Bezug auf untergeordnete Schreibfähigkeiten (z.B. Rechtschreibung) präsentiert. Die Formulierungen derselben wurden ebenfalls den Fragebögen beziehungsweise den Antworten der Lehrkräfte in der Piloterhebung entnommen. Vergleichbare Formulierungen finden sich jedoch auch in der Literatur/existierenden Bewertungsskalen (vgl. Hachmeister, 2019; Nussbaumer, 1991).

Die Teilnehmer*innen bewerteten die Bedeutung der einzelnen Kriterien für ihre Textbewertungen (d.h. deren Gewicht) auf einer sechsstufigen Ratingskala (1 = überhaupt nicht wichtig; 6 = sehr wichtig).

Abbildung 4: Ausschnitt der Fragebogenseite zur Gewichtung der Relevanz von Textbewertungskriterien

Tabelle 1: Übersicht der zu gewichtenden Dimensionen der Textproduktionskompetenz (TKP) und entsprechender Kriterien

Komponenten der TPK	**Kriterien**
Adressatenorientierung	● Berücksichtigung der Leser*innenperspektive (11) ● Darstellung der Ereignisse in sinnvoller Reihenfolge (12) ● Die vollständige Darstellung von Ereignissen (13)
Informationsmanagement	● Nachvollziehbarkeit der Geschehnisse durch genügend Informationen (1) ● Gabe möglichst vieler Informationen (2) ● Vermeidung von zu vielen Informationen (3)
Anwendung passender Textmuster(-aspekte) (hier: Erzählung)	● Vorhandensein einer Einleitung (4) ● Vorhandensein eines Schlusses (5) ● Nutzung des richtigen Tempus (6) ● Aufbau von Spannung im Verlauf des Textes (7) ● Einsatz wörtlicher Rede zur Erzeugung eines szenischen Charakters (8) ● Erzeugung emotionaler Momente (9) ● Unterstützung der Lesenden durch allgemeinen Aufbau des Textes (10)
Herstellung von Kohärenz	● Verwendung von Gliederungselementen (14) ● Stimmigkeit der Wortwahl (15) ● Stimmigkeit des Textes als Ganzes (16) ● Sinnvolle Beziehung zwischen den Sätzen eines Textes (17) ● Sinnvolle Beziehung zwischen den Wörtern eines Textes (18) ● Beachtung der Schreibaufgabe (19) ● Unterhaltsamkeit eines Textes (20)
Andere	● Vermeidung von Rechtschreibfehlern (21) ● Vermeidung von Grammatikfehlern (22) ● Die Textlänge (23) ● Der Wortschatz (24) ● Das Schriftbild (25)

Anmerkung: Die vorgestellten Kriterien wurden den Dimensionen der TPK theoriegeleitet zugeordnet. Aufgrund der theoretischen Überschneidungen zwischen diesen Kriterien kann die Kategorisierung gelegentlich unklar/uneindeutig sein. Andere Formulierungen zur Benennung der Kriterien wären möglich, die hier genutzten stammen aus bereits existierenden Bewertungsrastern und einer qualitativen Lehrkraftbefragung zu verwendeten Textbewertungskriterien.

5.4.4 Statistische Analysen

5.4.4.1 Grundzüge der Multi-Facetten-Rasch-Analyse

Zur Beantwortung der Forschungsfragen wurde eine Multi-Facetten-Rasch-Analyse (MFRA; Linacre, 2019) durchgeführt. Hierbei handelt es sich um ein gängiges Verfahren, wenn Aspekte des Verhaltens von Beurteiler*innen und/oder Beurteiler*innen-Effekte untersucht werden sollen. Im Kontext der Textbewertung und der Bewertung kreativer Produkte kam die MFRA bereits mehrfach zur Anwendung (vgl. Eckes, 2005; Eckes, 2008; He, Gou, Chien, Chen & Chang, 2013; Hodges et al., 2019; Primi, Silvia, Jauk & Benedek, 2019). Ein immenser Vorteil des Multi-Facetten-Rasch-Modells – vor allem gegenüber Analyseverfahren, die auf der klassischen Testtheorie basieren – besteht darin, dass im Rahmen der MFRA Effekte, welche von Beurteilenden ausgehen (z.B. Anwendung besonders strenger/milder Standards), systematisch berücksichtigt werden und nicht Teil der unsystematischen Fehlervarianz sind (Eckes, 2004). Die MFRA ermöglicht zudem auf Basis der systematischen Berücksichtigung von Beurteiler-Effekten eine Einschätzung der Strenge/Milde jeder einzelnen urteilenden Person und des damit verbundenen Messfehlers. Auf diese Weise ist es möglich, Aussagen zum Bewertungsverhalten jeder einzelnen Person zu treffen. Ein weiterer Vorteil des Modells besteht darin, dass es – ebenso wie andere Rasch-Modelle – bei der Modellschätzung die Personenfähigkeit (hier deren Urteilsverhalten) und die Itemschwierigkeit (hier die Textqualität/das Gewicht der Kriterien) getrennt voneinander berücksichtigt. Eine Konfundierung dieser beiden Variablen ist damit nicht zu befürchten (Eckes, 2004).

Die MFRA basiert auf dem von Andrich (1978) vorgeschlagenen Modell der Ratingskala für polytome Antworten. Sie erlaubt die gleichzeitige Modellierung psychometrischer Merkmale einer Produktbewertung (z.B. der Textqualität) sowie verschiedener Aspekte des Verhaltens von Beurteiler*innen, Beurteiler*innen-Effekte und Eigenschaften der Beurteiler*innen (Wu, 2017). Die MFRA kann zur Bestimmung der Objektivität eines Ratings verwendet werden. Im Hinblick auf die aufgeworfenen Forschungsfragen erweist sich das Verfahren demnach als besonders geeignet.

Es wurden zwei verschiedene Modelle berechnet. Das erste Modell (Modell A) analysierte die globalen Textqualitätsbewertungen, während das zweite Modell (Modell B) die Gewichtung der Bewertungskriterien analysierte.

Die grundlegende Gleichung für beide Modelle ist gegeben durch:

$$log\left(\frac{p_{v,j,k}}{p_{v,j,k-1}}\right) = \vartheta v - \alpha j - \tau k$$

Spezifikation Modell A: In Modell A steht $p_{v,j,k}$ für die Wahrscheinlichkeit, mit der ein bestimmter Text (v), bewertet durch eine bestimmte Person (j), die Note k (z.B. „3“ auf der 6-stufigen Bewertungsskala) erhält. In ähnlicher Weise steht $p_{v,j,k-1}$ für die Wahrscheinlichkeit, die Note k-1 zu erhalten (z.B. „2“ im obigen Beispiel). Das Verhältnis der beiden (Odds Ratio) stellt die Wahrscheinlichkeit dar, mit der die bewertende Person (j) einen bestimmten Wert k (z.B. „3“) vergibt, verglichen mit der Wahrscheinlichkeit, mit der die bewertende Person einen Wert von k-1 (z.B. „2“) vergibt. In diesem Modell hängt der Wert des Odds Ratio von drei Parametern ab: dem Qualitätsparameter (ϑv) eines bestimmten Textes, dem Strengeparameter einer bewertenden Person (αj), und dem Schwellwertparameter (τk), der den Übergangspunkt darstellt, an dem die Wahrscheinlichkeit, dass ein Text in jede der beiden angrenzenden Kategorien fällt, 50% beträgt (vorausgesetzt, er fällt überhaupt in eine dieser beiden Kategorien; Eckes, 2009). Diese Parameter sind entscheidend für die empirische Anordnung der Kategorien der Ratingskala und damit für die Gesamtfunktionalität der Ratingskala. In einem Ratingskalenmodell (Andrich, 1978) werden die Schwellenparameter über alle bewertenden Personen hinweg gleichgesetzt, d.h. die Ratingskala funktioniert für alle bewertenden Personen gleich. Durch das Hinzufügen eines Gruppenzugehörigkeitsparameters (γg) können gruppenspezifische Parameter der beurteilenden Personen geschätzt werden. Ein sehr nützliches Merkmal des Ratingskalenmodells ist, dass alle Parameter des Modells auf der gleichen Logit-Skala kalibriert und somit direkt vergleichbar sind.

Im zweiten Modell (Modell B) wurde das gleiche Modell an die Daten angepasst. Diesmal wird jedoch ein bestimmtes Kriterium (v) aus der Liste in Anhang A von einer bestimmten Person (j) bewertet. Daher hängt der Wert der Odds Ratio in Modell B von folgenden Parametern ab: dem Qualitätsparameter (ϑv) für ein bestimmtes Kriterium, welches die Bedeutung widerspiegelt, die die bewertenden Personen dem Kriterium beimessen (d.h. die Gewichtung des Kriteriums für den Textbewertungsprozess); dem Strengeparameter für eine bewertende Person (αj), welches die globale Tendenz der jeweiligen Person widerspiegelt, den Kriterien allgemein

mehr oder weniger Relevanz beizumessen; und dem Schwellenwert-Parameter (τk), wie in Modell A beschrieben. Auch hier kann ein Gruppenzugehörigkeitsparameter (γg) hinzugefügt werden, um gruppenspezifische Parameter zu schätzen.

5.4.4.2 Kennwerte der Multi-Facetten-Rasch-Analyse

Modellpassung: Der globale Fit des Multifacetten-Rasch-Modells wird anhand der Residuen zwischen der beobachteten und der vom Modell vorhergesagten Antworten berechnet. Von einer guten Modellanpassung kann nach Linacre (2012) dann ausgegangen werden, wenn 5% (oder weniger) der standardisierten Residuen gleich oder größer 2 und 1% (oder weniger) der standardisierten Residuen gleich oder größer 3 sind.

Darüber hinaus werden im Rahmen der MFRA verschiedene Kennwerte berechnet, die Rückschlüsse auf die Eigenschaften der untersuchten Variablen im Modell (hier *Facetten* genannt) und mögliche Beurteiler*innen-Effekte zulassen. Kennwerte, die für die Beantwortung der in dieser Studie gestellten Forschungsfragen relevant sind, werden im Folgenden kurz beschrieben:

Strengemaße geben Aufschluss über die Bewertungsmaßstäbe der Beurteiler*innen (die Beurteiler*innenstrenge; Eckes, 2004). Während positive Strengemaße (positive Logit-Werte) die Tendenz der Bewertenden zur Vergabe eher niedriger Bewertungen anzeigen, implizieren negative Strengemaße (negative Logit-Werte) die Tendenz der Bewertenden zur Vergabe höherer Bewertungen.

Der Chi-Quadrat-Test wird verwendet, um zu testen, ob die Elemente einer Facette (z.B. die Bewertenden einer Gruppe) aus einer homogenen Population stammen und sich somit bestimmte Parameter (z.B. die Beurteiler*innenstrenge) über alle Elemente einer Facette hinweg fixieren lassen (Linacre, 2003; Linacre, 2012).

Die Separationsreliabilität gibt Auskunft darüber, inwieweit sich einzelne Elemente einer Facette (z.B. Bewertende innerhalb einer Gruppe, oder Texte hinsichtlich ihrer Qualität) voneinander unterscheiden. Je höher der Wert der Separationsreliabilität ist, desto deutlicher sind die Unterschiede zwischen den einzelnen Elementen einer Facette.

Der Index der Klassenseparation gibt an, in wie viele unterschiedliche Klassen die Elemente einer einzelnen Facette eingeteilt werden können (Wright & Masters, 2002). Elemente innerhalb einer gebildeten Klasse sind einander ähnlich, zwischen den Elementen verschiedener Klassen bestehen dahingegen deutliche Unterschiede. Ergibt sich für eine Facette lediglich eine

Klasse, zeigt dieses Ergebnis an, dass die Unterschiede zwischen den einzelnen Elementen einer Facette gering und damit zu vernachlässigen sind (d.h. beispielsweise: die Bewertungsmaßstäbe der Beurteiler*innen innerhalb einer Gruppe wären miteinander vergleichbar). Gibt der Index mehr als eine Klasse an, ist dies ein Hinweis auf inhomogene Facetten (d.h. beispielsweise: die Bewertungsmaßstäbe der Beurteiler*innen innerhalb einer Gruppe unterscheiden sich voneinander. Es gibt mindestens zwei Klassen von Bewertenden, die jeweils unterschiedliche Maßstäbe anlegen).

Mean-Square-Fehlerstatistiken geben Auskunft darüber, wie gut die einzelnen Elemente einer Facette den Erwartungen des Modells entsprechen. Innerhalb des MFRA werden diese Statistiken zur Analyse der Urteilskonsistenz einzelner Beurteiler*innen genutzt (Linacre, 2002). Bei Ratern, die eine zufriedenstellende Modellanpassung aufweisen, wird angenommen, dass sie konsistente Bewertungen (in Bezug zur Erwartung des Modells) abgeben. Zur Bestimmung der Modellpassung werden zwei verschiedene Anpassungsmaße berechnet: Infit- und Outfit-Statistiken. Diese sind vergleichbar mit gleichnamigen Kennwerten, die im Rahmen von Itemanalysen auf der Basis des Rasch-Modells ermittelt werden, beziehen sich hier aber auf die Modellpassung von Bewertenden bzw. ihrer Urteile. Die Infit-Statistik beschreibt das Verhalten von Bewerter*innen (d.h. konsistent vs. inkonsistent) bei der Bewertung von Produkten, deren Qualität (in Logits) nahe am Strengemaß der/des jeweiligen Bewertenden liegt (maximale +/- 0,5 Logits) und damit nahe des Ankers der/des jeweiligen Bewertenden für „durchschnittliche“ Qualität. Umgekehrt beschreibt die Outfit-Statistik das Verhalten von Bewerter*innen bei der Beurteilung sehr schlechter oder sehr guter Produkte im Vergleich zu ihrem individuellen Anker (Strengemaß). Die Werte für beide Statistiken reichen von 0 bis ∞. Anpassungswerte von 1 repräsentieren eine perfekte Modellanpassung, während Werte größer 1 auf eine größere Variation in den Bewertungen hinweisen als vom Modell erwartet (Bewerter*innen mit einem Anpassungswert von 1,25 weisen 25% mehr Varianz in ihren Bewertungen auf als erwartet). Im Gegensatz dazu weisen Werte kleiner 1 auf eine geringere Variation als erwartet hin (Beurteiler*innen mit einem Anpassungswert von 0,70 weisen 30% weniger Varianz in ihren Bewertungen auf als erwartet). Abhängig von der Richtung, in die Bewertende vom Modell abweichen, können diese entweder als *überangepasst* (weniger Varianz in den Bewertungen als erwartet und daher zu leicht vorhersagbar) oder als *unterangepasst* (mehr Varianz in ihren Bewertungen als erwartet und daher zu wenig vorhersagbar) klassifiziert werden.

5.5 Ergebnisse

5.5.1 Urteile zur globalen Textqualität (Model A)

5.5.1.1 Facette der ‘Textqualität’

Zur Beurteilung der Textqualität nutzen die Bewertenden nahezu alle Punktwerte der sechsstufigen Bewertungsskala aus. Lediglich der Wert 1 (= sehr schlechte Qualität) wurde nicht vergeben. Der Chi-Quadrat-Test zeigte, dass sich die präsentierten 20 Texte hinsichtlich ihrer Qualität signifikant voneinander unterscheiden ($\chi^2_{(19)}$=3728,3; p<0,002); zwischen den Texten bestehen deutliche Qualitätsunterschiede (Separationsreliabilität = 1,00). Die Textqualität der 20 Schüler*innentexte variiert damit (wie intendiert) über ein breites Spektrum; sowohl sehr gute als auch schlechte Texte sind in der Stichprobe enthalten (siehe Facettenraum, Anhang B).

5.5.1.2 Strenge der Beurteiler*innen bei der Bewertung der Textqualität (Forschungsfrage 1)

Zur Beantwortung der Frage danach, wie streng die Textbewertungen von Lehrkräften im Vergleich zu den Bewertungen anderer (potentieller) Experten und Novizen sind, wurden die (durchschnittlichen) Strengemaße aller untersuchten Gruppen miteinander verglichen.

Insgesamt verteilten sich die Strengemaße der Bewerter*innen zwischen -2,10 (mildeste Person) und +2,07 (strengste Person) auf der Logit-Skala (siehe Facettenraum, Anhang A). Signifikante Gruppenunterschiede hinsichtlich der (mittleren) Strenge konnten nicht gefunden werden ($\chi^2_{(3)}$=3,9; p=0,27). Die Gruppen sind demnach hinsichtlich ihrer Strenge bei der Bewertung der Textqualität miteinander vergleichbar; keine der Gruppen unterscheidet sich in ihrem durchschnittlichen Strengemaß von der Gruppe der Novizen.

Allerdings zeigten sich innerhalb der einzelnen Gruppen signifikante Unterschiede im Hinblick auf die Strengemaße der einzelnen Gruppenmitglieder. (Der Chi-Quadrat-Test fällt für alle Gruppen signifikant aus; siehe Tabelle 2.) Das Ausmaß der vorgefundenen Unterschiede ist innerhalb der verschiedenen Gruppen vergleichbar, denn sowohl die Separationsreliabilität als auch der Index der Klassenseparation weisen vergleichbare Werte für die untersuchten Gruppen auf (siehe Tabelle 2).

Tabelle 2. Beurteilerstrenge und Separationsstatistiken bei der globalen Textbewertung: Stichprobengröße, M (SD) des durchschnittlichen Strengemaßes, Chi-Quadrat-Test χ^2 (df), Separationsreliabilität, Index der Klassenseparation (Anzahl der Subgruppen).

	Lehrkräfte	Wissenschaftler*innen	Lektor*innen & Autor*innen	Novizen
Stichprobengröße	51	24	31	69
Strengemaß	-0,03 (0,04)	0,06 (0,06)	0,03 (0,05)	-0,05 (0,03)
Chi-Quadrat-Test	325,8 (50)*	140,0 (23)*	273,9 (30)*	530,2 (68)*
Separationsreliabilität	0,85	0,83	0,89	0,88
Klassenseparation	2,38	2,25	2,90	2,69

* $p<0,05$

5.5.1.3 Konsistenz der Bewertungen der globalen Textqualität (Forschungsfrage 2)

Im Rahmen der MFRA lässt sich die Konsistenz/Inkonsistenz der Bewertungen einer Person anhand ihrer Modellpassung bestimmen. Zur Beantwortung der Frage danach, wie konsistent die Textbewertungen von Lehrkräften im Vergleich zu anderen (potentiellen) Experten und Novizen sind, wurde der Prozentsatz von Bewertenden mit guter sowie schlechter Modellpassung zwischen den Gruppen verglichen. Der Bereich der Fehlerstatistiken, die eine gute Modellanpassung anzeigen, wurde entlang Linacre und Wright (1994) auf 0,4 bis 1,2 festgelegt. In einem ersten Schritt wurden entlang dieser Werte alle Bewertenden anhand ihrer individuellen Mean-Square-Fehlerstatistiken als passend oder unpassend zum Modell kategorisiert. In einem zweiten Schritt wurde dann der Prozentsatz konsistent beziehungsweise nichtkonsistent urteilender Bewerter*innen innerhalb jeder Gruppe bestimmt (siehe Tabelle 3). Die ermittelten Prozentsätze der verschiedenen Gruppen wurden dann miteinander verglichen.

Tabelle 3. Anteil der konsistent urteilenden Bewerter*innen im Gruppenvergleich.

	Lehrkräfte	Wissenschaftler*innen	Lektor*innen & Autor*innen	Novizen
Infit	76,0	87,5	68,0	66,5
Outfit	78,0	87,5	68,0	69,0

Anmerkungen: Infit- Statistiken beschreiben Bewertungen von Texten mit „durchschnittlicher“ Qualität; Outfit-Statistiken beschreiben Bewertungen von „extrem schlechten“ oder „extrem guten“ Texten – jeweils im Verhältnis zum individuellen Anker (Strengemaß) der Bewertenden.

Im Hinblick auf die Prozentsätze modellkonformer, d.h. konsistent urteilender Bewerter*innen ergaben sich keine signifikanten Gruppenunterschiede, weder im Hinblick auf die Infit-Statistik ($\chi^2_{(3)}$=4,03; p=0,26) noch im Hinblick auf die Outfit-Statistik ($\chi^2_{(3)}$ =3,70; p=0,30). Alle Gruppen sind somit im Hinblick auf ihre Konsistenz, d.h. die Übereinstimmung der Bewertenden bei der globalen Textbeurteilung, miteinander vergleichbar. Weder die Gruppe der Lehrkräfte noch eine der beiden Expertengruppen ist in ihren Bewertungen konsistenter als die Gruppe der Novizen.

Deskriptiv betrachtet, findet sich in der Gruppe der Lehrkräfte der zweithöchste Anteil konsistenter Bewerter*innen – dies gilt gleichermaßen für die Infit- und die Outfit-Statistik. Allerdings wichen mehr als 20% der Lehrkräfte von den Erwartungen des Modells ab und gaben somit Bewertungen ab, die nicht konsistent mit denen anderer Lehrkräfte waren. Wissenschaftler*innen waren in ihren Bewertungen konsistenter als Lehrkräfte; Der Anteil inkonsistent urteilender Wissenschaftler*innen betrug lediglich 12,5% – etwa halb so viel wie unter den Lehrkräften. In der Gruppe der Lektor*innen und Autor*innen ebenso wie in der Gruppe der Novizen, war der Anteil an Bewertenden, deren Textbeurteilungen nicht konsistent mit denen anderer Gruppenmitglieder sind, am höchsten.

5.5.1.4 Modellpassung

Modell A ist gut angepasst. Nur 3,9% der standardisierten Residuen sind gleich oder größer 2 und nur 0,6% der standardisierten Residuen sind gleich oder größer 3. Die Daten können dementsprechend gut durch das Modell beschrieben werden.

5.5.2 Relevanz der Bewertungskriterien für die globalen Textbeurteilungen (Modell B)

5.5.2.1 Facette Gewichtung/Relevanz der Kriterien

Zur Gewichtung der vorgegebenen Kriterien verwendeten die Bewerter*innen nur einige Skalenwerte der sechsstufigen Bewertungsskala. Die Skalenwerte 1 (= „völlig irrelevant“) und 2 (= „irrelevant“) wurden nicht verwendet (siehe Facettenraum, Anhang B). Dies gilt sowohl in Bezug auf Kriterien, die sich auf die Dimensionen der TPK beziehen als auch in Bezug auf die eher hierarchieniedrigen Kriterien (z.B. Rechtschreibung). Der Chi-Quadrat-Test zeigt, dass sich die 25 vorgegebenen Kriterien hinsichtlich der ihnen zugeschriebenen Relevanz/Gewichtung signifikant voneinander unterscheiden ($\chi^2_{(24)}$=1618,2; p< 0,001) und die vorgefundenen Unterschiede als deutlich beschrieben werden können (Separationsreliabilität = 0,98).

5.5.2.2 Gruppenunterschiede in der Gewichtung der Bewertungskriterien

Insgesamt variierten die Strengeparameter der Bewerter*innen von -1,7 (Gewichtung der Kriterien insgesamt als eher wichtig) bis +1,1 (Gewichtung der Kriterien insgesamt als weniger wichtig) auf der Logit-Skala (siehe Facettenraum, Anhang B).

Die Unterschiede zwischen den Gruppen in Bezug auf die Strengeparameter, d.h. in Bezug auf das Gesamtgewicht, welches dem gegebenen Satz an Kriterien zugeschrieben wird (siehe Tabelle 3), erreichten ein signifikantes Niveau ($\chi^2_{(3)}$=21,6; p<0,001).

Für die Gruppe der Lehrkräfte aber auch für die Gruppe der Lektor*innen und Autor*innen gilt, dass diese die vorgegebenen Textbewertungskriterien insgesamt als eher wichtig für die Beurteilung der Textqualität einstufen; die durchschnittlichen Strengeparameter beider Gruppen sind negativ. Im Gegensatz dazu sind die durchschnittlichen Strengeparameter der Wissenschaftler*innen und Novizen positiv, was anzeigt, dass diese dem vorgegebenen Satz von Kriterien weniger Gewicht beimessen.

Der Chi-Quadrat-Test zeigt (ebenso wie in Model A), dass in Bezug auf die Gewichtung der Bewertungskriterien signifikante Unterschiede innerhalb aller Gruppen bestehen. Die Separationsreliabilität nimmt in allen Gruppen einen hohen Wert an (siehe Tabelle 4), folglich unterscheiden sich die Bewerter*innen innerhalb ihrer jeweiligen Gruppen deutlich voneinander hinsichtlich der Gewichte, welche sie den Kriterien zuschreiben. Zusätzlich zeigt der Index der Klassenseparation an, dass in allen Gruppen mehrere Klassen, d.h. Subgruppen mit unter-

schiedlichen Bewertungsmaßstäben, existieren (siehe Tabelle 4). Allerdings zeigen die Wissenschaftler*innen entlang der Separationsstatistiken bei der Gewichtung der Kriterien untereinander mehr Ähnlichkeiten als Personen der anderen Gruppen.

Tabelle 4. Beurteilerstrenge und Separationsstatistiken bei der Gewichtung der Kriterien: Stichprobengröße, M (SD) des durchschnittlichen Strengeparameters, Chi-Quadrat-Test χ^2 (df), Separationsreliabilität, Index der Klassenseparation (Anzahl der Klassen/Subgruppen).

	Lehrkräfte	Wissenschaftler*innen	Lektor*innen & Autor*innen	Novizen
Stichprobengröße	45	20	30	58
durchschnittlicher Strengeparameter	-0,09 (0,03)	0,11 (0,04)	-0,07 (0,04)	0,05 (0,03)
Chi-Quadrat-Test	180,8 (44)*	61,6 (19)*	179,9 (29)*	297,6 (57)*
Separationsreliabilität	0,80	0,69	0,84	0,83
Klassenseparation	1,99	1,48	2,32	2,17

Anmerkung: Aus rechnerischen Gründen repräsentieren kleinere Einheiten die Tendenz zur Vergabe höherer Gewichte; * $p<0,05$

5.5.2.3 Konsistenz in den Gewichtungen der Textbewertungskriterien (Forschungsfrage 3)

Zur Beantwortung der Forschungsfrage, wie konsistent Lehrkräfte in den Gewichtungen der Textbewertungskriterien im Vergleich zu anderen Experten und Novizen sind, wurde entlang des gleichen Verfahrens wie oben (siehe Forschungsfrage 2) der Prozentsatz von Bewertenden mit guter/schlechter Modellpassung pro Gruppe bestimmt und anschließend zwischen den Gruppen verglichen (siehe Tabelle 5).

Tabelle 5. Anteil der Bewerter*innen, die konsistente Gewichtungen der Kriterien vornehmen, im Gruppenvergleich

	Lehrkräfte	Wissenschaftler*innen	Lektor*innen & Autor*innen	Novizen
Infit	80.0	75.0	50.0	69.0
Outfit	82.0	75.0	57.0	72.0

Anmerkungen: Infit- Statistiken beschreiben Gewichtungen von Kriterien mit „durchschnittlicher“ Relevanz; Outfit-Statistiken beschreiben Gewichtungen von Kriterien mit „sehr geringer“ oder „sehr hoher“ Relevanz – jeweils im Verhältnis zum individuellen Anker (Strengemaß) der Bewertenden.

Für die Outfit-Statistik erreichten die Unterschiede in der Konsistenz zwischen den Gruppen kein signifikantes Niveau ($\chi^2_{(3)}$=5,97; p=0,11). Für die Infit-Statistik wurden jedoch signifikante Unterschiede hinsichtlich der Konsistenz zwischen den Gruppen gefunden ($\chi^2_{(3)}$=8,0; p=0,046). Hieraus lässt sich schlussfolgern, dass alle Gruppen ein vergleichbares Maß an Konsistenz bei der Gewichtung besonders wichtiger und besonders unwichtiger Kriterien (im Verhältnis zum jeweiligen Anker der Bewertenden) aufweisen. Die Konsistenz innerhalb der Gruppen unterscheidet sich jedoch bei der Gewichtung von Kriterien, die von „durchschnittlicher“ Relevanz (auch hier im Verhältnis zum jeweiligen Anker der Bewertenden) sind und deren Einordnung damit weniger eindeutig ist. In der Gruppe der Lektor*innen und Autor*innen ist der Anteil der Bewerter*innen, die bei der Gewichtung von Kriterien mit „durchschnittlicher“ Relevanz konsistente Bewertungen abgeben, signifikant geringer als in den anderen untersuchten Gruppen (mit standardisierten Residuen < 2,0). Der Anteil der Personen, der die Relevanz der Bewertungskriterien konsistent gewichtet (Personen mit entsprechender Modellpassung), ist in der Gruppe der Lehrkräfte am höchsten, gefolgt von der Gruppe der Wissenschaftler*innen und der der Novizen. Diese drei Gruppen unterscheiden sich statistisch nicht voneinander.

5.5.2.4 Modellpassung

Modell B ist gut angepasst. Nur 3,1% der standardisierten Residuen sind gleich oder größer 2 und 0,9% der standardisierten Residuen sind gleich oder größer 3. Auch in diesem Fall (Modell B) können die Daten gut durch das Modell beschrieben werden.

5.5.2.5 Gewichtung der einzelnen Kriterien (Forschungsfrage 4)

Mit Blick auf die Gewichtung der einzelnen Kriterien (mittlere Logit-Werte) zeigen sich auffällige Unterschiede zwischen den Gruppen (siehe Abbildung 5).

Lehrkräfte schreiben den folgenden Aspekten eine höhere Relevanz zu als Novizen:

- Kriterium 1 (*Informationsmanagement*) – „Nachvollziehbarkeit der Geschehnisse durch genügend Informationen“
- allen Kriterien der *Textmusterdimension* (außer Kriterium 10 – „Unterstützung der Lesenden durch allgemeinen Aufbau des Textes“)
- Kriterium 13 (*Adressatenorientierung*) – „vollständige Darstellung von Ereignissen“
- drei Aspekten der Dimension *Kohärenz*: Kriterium 14 – „Verwendung von Gliederungselementen“, Kriterium 15 – „Stimmigkeit der Wortwahl“ und Kriterium 19 – „Beachtung der Schreibaufgabe“

„Rechtschreibfehler“ (Kriterium 21) werden von Lehrkräften als weniger relevant eingestuft als von Novizen.

Im Vergleich aller Gruppen zeigt sich, dass

- für Lehrkräfte die Kriterien (4) „Vorhandensein einer Einleitung“, (5) „Vorhandensein eines Schlusses“ und (19) „Beachtung der Schreibaufgabe“ bei der Bewertung der Textqualität wichtiger sind als für alle anderen Gruppen.
- für Wissenschaftler*innen sind die Kriterien (10) „Unterstützung der Lesenden durch allgemeinen Aufbau des Textes“ und (11) „Berücksichtigung der Leser*innenperspektive“ bei der Bewertung der Textqualität wichtiger als für alle anderen Gruppen.
- für Lektor*innen und Autor*innen ist das Kriterium 20 – „Unterhaltsamkeit des Textes“ (d.h. die Textfunktion) wichtiger als für alle anderen Gruppen.
- für Novizen ist keines der vorgestellten Kriterien wichtiger als für die anderen Gruppen.

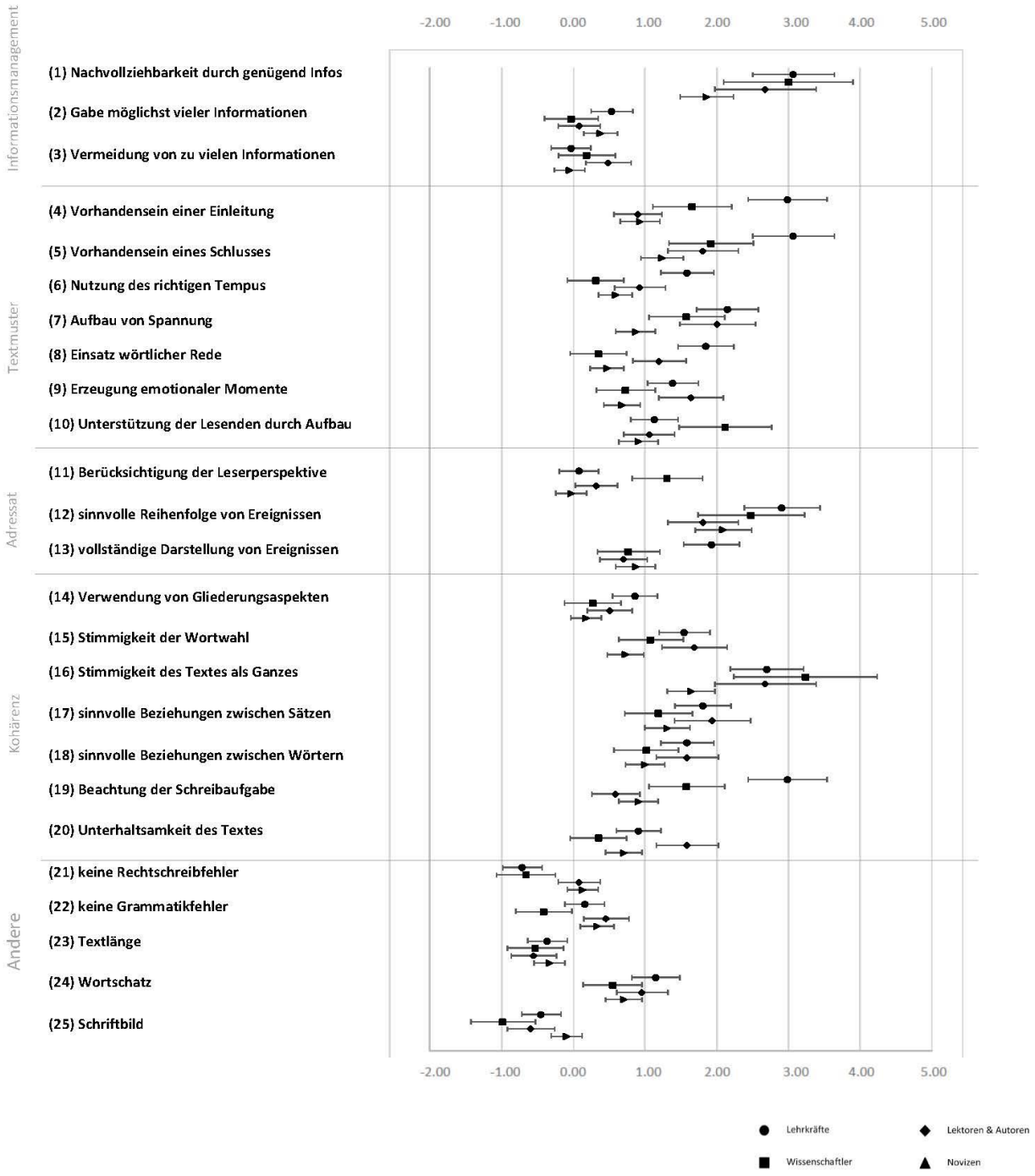


Abbildung 5: Gewichtungen der einzelnen Textbewertungskriterien im Gruppenvergleich. Anmerkung: Nicht überlappende Konfidenzintervalle markieren signifikante Gruppenunterschiede.

Unter Nutzung des 75. Perzentils aller Logit-Werte als Maßstab für besonders wichtige Kriterien zeigt sich, dass zwei der vorgegebenen 25 Kriterien für beide Expertengruppen und auch für Lehrkräfte von besonderer Relevanz sind. Hierbei handelt es sich um Kriterium 1 „Nachvollziehbarkeit der Geschehnisse durch genügend Informationen“ und Kriterium 16 „Stimmigkeit des Textes als Ganzes“. Kein einziges Kriterium ist für Novizen von besonderer Relevanz.

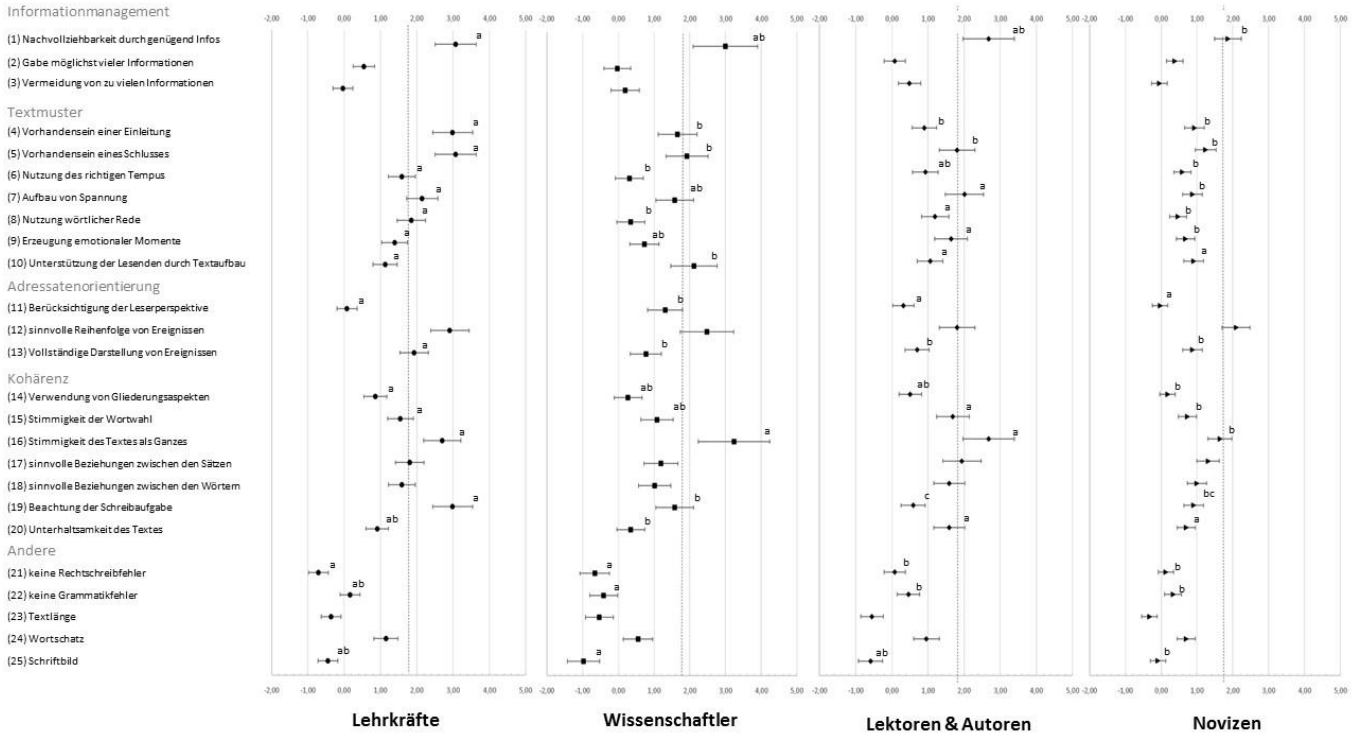


Abbildung 6: Relevanz der einzelnen Textbewertungskriterien gemessen am 75. Perzentil aller Mittelwerte. Werte auf der rechten Seite der gestrichelten Linie zeigen an, dass die Gruppenmittelwerte über dem 75. Perzentil aller Mittelwerte liegen. ^{abc} gleiche Indizes bezeichnen homogene Subgruppen bei $\alpha = 5\%$. Weitere Informationen zu den Kriterien finden Sie in Anhang B.

5.6 Zusammenfassung und Diskussion der Ergebnisse

Die Hauptfragestellung dieser Studie (Studie 1) lautete: Sind Lehrkräfte Experten bei der Bewertung von (Schüler*innen-)Texten? Entlang des Experten-Novizen-Paradigmas wurden drei Kriterien identifiziert, anhand derer sich Experten von Novizen abgrenzen lassen: 1. Die Strenge in den Bewertungen von Experten (im Gegensatz zur Milde der Bewertungen durch Novizen), 2. die Konsistenz zwischen den Bewertungen verschiedener Experten (im Gegensatz zur Inkonsistenz zwischen den Bewertungen verschiedener Novizen) und 3. die Nutzung valider Kriterien beziehungsweise deren konsistente Gewichtung bei der Bewertung von Texten durch Experten (im Gegensatz zu Novizen, die hierin weniger konsistent sind und wichtige Kriterien nicht als solche erkennen). Anhand der vorliegenden Daten zeigt ein Vergleich von Lehrkräften und Novizen, dass diese sich hinsichtlich 1) der Strenge ihrer Bewertungen nicht voneinander unterscheiden; Lehrkräfte neigen ebenso wie Novizen zu relativ milden Bewertungen. Außerdem finden sich im Hinblick auf 2) die Konsistenz der Textbewertungen keine Gruppenunterschiede zwischen Lehrkräften und Novizen und auch mit Blick auf 3) die Konsistenz der Gewichtungen von Bewertungskriterien unterscheidet sich die Gruppe der Lehrkräfte nicht von der der Novizen.

In einigen Punkten ließen sich jedoch signifikante Unterschiede zwischen Novizen und Lehrkräften finden: Zum einen ergaben sich signifikante Unterschiede zwischen den beiden Gruppen im Hinblick auf die Relevanz, welche diese den vorgegebenen Kriterien im Durchschnitt zuschrieben. Lehrkräfte räumen dem vorgegebenen Satz an Kriterien insgesamt ein größeres Gewicht für die Beurteilung der Textqualität ein als Novizen. Außerdem unterscheiden sich die beiden Gruppen mit Blick auf einzelne Bewertungskriterien signifikant voneinander. Für Lehrkräfte sind 11 der 25 vorgegebenen Kriterien deutlich wichtiger als für Novizen. Diese besonders wichtigen Kriterien beziehen sich vor allem auf den Einsatz geeigneter Textmustermerkmale und die Herstellung von Kohärenz. Andererseits zeigten sich jedoch bei 13 der 25 Kriterien keine signifikanten Unterschiede zwischen Lehrkräften und Novizen. Lediglich ein Kriterium (Kriterium 21 – Rechtschreibung) wurde von Novizen als deutlich wichtiger eingestuft als von Lehrkräften.

Zusammengenommen zeigen die Daten bei der Beurteilung der Qualität von (Schüler*innen-)Texten somit mehr Gemeinsamkeiten als Unterschiede zwischen Lehrkräften und Novizen. Die beiden Gruppen unterscheiden sich nicht hinsichtlich ihrer Strenge oder Konsistenz bei der Textbewertung (sowie der Gewichtung der hierfür verwendeten Kriterien). Es lassen sich nur wenige lehrkraftspezifische Merkmale finden.

Wie in Kapitel 3 dieser Arbeit dargelegt, ist dies insofern problematisch, als dass die Textproduktionskompetenz eine grundlegende Schlüsselfähigkeit zur Teilhabe in der modernen Gesellschaft darstellt und Lehrerinnen und Lehrer daher in der Lage sein sollten, Schüler*innen bei der Entwicklung der Textproduktionskompetenz (zumindest hin zu einem gewissen Grundniveau) zu unterstützen. Alle in dieser Studie vorgegebenen Textbeurteilungskriterien (mit Ausnahme der letzten fünf eher grundlegenden Kriterien wie z.B. der Rechtschreibung) stehen in Verbindung/Einklang mit nationaler (und auch internationaler) Literatur zu hierarchiehohen Komponenten der Textproduktionskompetenz (oder auch Schreibkompetenz; Augst, Disselhoff, Henrich, Pohl & Völzing, 2007; Crossley & McNamara, 2016; Dam-Jensen & Heine, 2013; Hayes & Flower, 1980; Hayes, 2012; Hennes et al., 2018; Kellogg, 2008; Knopp et al., 2014; MacArthur & Graham, 2016). Trotzdem messen Lehrkräfte dem Großteil dieser Kriterien nicht mehr Bedeutung zu als Novizen (obwohl Lehrkräfte die Relevanz der Kriterien insgesamt höher einordnen). Lehrkräfte schreiben weniger als der Hälfte der vorgegebenen Kriterien ein höheres Gewicht zu als Novizen. Unter den von Lehrkräften nicht ausreichend beachteten Kriterien befinden sich wichtige Aspekte wie die „Unterstützung der Lesenden durch allgemeinen Aufbau des Textes“ oder die „Sinnvolle Beziehung zwischen den Sätze eines Textes“ (beiden Kriterien wird nur eine „durchschnittliche“ Relevanz zugeschrieben). Als spontane Schlussfolgerung aus den Ergebnissen ließe sich ableiten, dass Lehrkräfte besser ausgebildet werden sollten, damit sie wichtige Kriterien zur Textbewertung anwenden und so verlässliche und konsistente (Experten-)Bewertungen abgeben können.

Ähnliche Schlussfolgerungen wurden in der Literatur schon häufig gezogen (Beck et al., 2018; Hodges et al., 2019; Penner-Williams, Smith & Gartin, 2009). Die vorliegende Studie geht jedoch über diese Argumentation hinaus, indem sie die Frage danach stellt, ob Experten der Textbewertung existieren, wenn Lehrkräfte nicht als solche gelten und wenn ja, wer diese Experten dann sind. Auf Basis der vorliegenden Daten muss mit Blick auf diese Frage jedoch festgestellt werden, dass sich keine der in dieser Studie untersuchten Expertengruppen (d.h. weder Wissenschaftler*innen noch Lektor*innen und Autor*innen) in (1) ihrer Strenge und (2) der Konsistenz bei der Beurteilung der Textqualität signifikant von der Gruppe der Novizen unterscheidet. Und auch im Hinblick auf (3) die Konsistenz bei der Gewichtung wichtiger Textbewertungskriterien unterscheiden sich die Wissenschaftler*innen nicht von den Novizen; die Gruppe der Lektor*innen und Autor*innen ist hier sogar teilweise weniger konsistent im Vergleich zur Gruppe der Novizen.

Außerdem zeigt sich bezogen auf die globale Gewichtung der vorgegebenen Textbewertungskriterien ein besonders interessanter Befund: Die Gruppe der Wissenschaftler*innen liegt in Bezug auf ihre durchschnittliche Gewichtung der in der vorliegenden Studie vorgegebenen Kriterien näher an den Novizen als an den Lehrkräften oder Lektor*innen und Autor*innen. Wissenschaftler*innen (ebenso wie Novizen) schreiben den Kriterien eher weniger Gewicht/Relevanz zu; Lehrkräfte und Lektor*innen und Autor*innen geben den Kriterien insgesamt ein größeres Gewicht. Dieser Befund ist vor allem deshalb interessant, da die Gruppe der Wissenschaftler*innen aufgrund ihres Engagements in der Forschung zur Textbewertung als Expertengruppe ausgewählt wurde. Es bestand die Annahme, dass gerade diese Gruppe über explizites Wissen zu relevanten Textbewertungskriterien und deren Bedeutung verfügt. Gemäß den Logit-Werten in Abbildung 5 gewichten Wissenschaftler*innen jedoch nur wenige Kriterien (Kriterium 10: „Unterstützung der Lesenden durch allgemeinen Aufbau des Textes“ und Kriterium 11: „Berücksichtigung der Leser*innenperspektive“) besonders stark.

Eine weitere Möglichkeit, die vorliegenden Daten zu interpretieren, stellt die Betrachtung der vorgefundenen Differenzen zwischen Novizen auf der einen Seite und den Lehrkräften sowie Expertengruppen auf der anderen Seite dar. Anhand des Maßstabs für besonders wichtige Kriterien (hier das 75. Perzentil aller Logit-Werte; Abbildung 6) lässt sich zeigen, dass zwei der 25 vorgegebenen Kriterien für alle Experten (Lehrkräfte eingeschlossen) relevant sind, nicht aber für Novizen. Hierbei handelt es sich um zwei durchaus wichtige Kriterien, die theoretischen Überlegungen zufolge maßgeblich zur Textqualität beitragen (Kriterium 1 „Bereitstellung von genügend Informationen, um das Geschehen eines Textes nachvollziehbar zu machen“ und Kriterium 16 „Stimmigkeit des Textes als Ganzes“). Insgesamt bleibt jedoch festzustellen, dass zwischen den Expertengruppen keine zufriedenstellende Übereinstimmung bezüglich der Relevanz der in der Literatur verwendeten Kriterien besteht. Drei von 25 Kriterien sind nur für Lehrkräfte von besonderer Relevanz und haben für die anderen Gruppen keine besondere Bedeutung (Kriterium 7 „Aufbau von Spannung im Verlauf des Textes“, Kriterium 1 „Nachvollziehbarkeit der Geschehnisse durch genügend Informationen“, Kriterium 19 „Beachtung der Schreibaufgabe“). Ein Kriterium ist besonders wichtig für Wissenschaftler*innen (Kriterium 10 „Unterstützung der Lesenden durch allgemeinen Aufbau des Textes“).

Als mögliche Erklärung für die vorgefundenen Unterschiede und gleichzeitig als Kritikpunkt an dieser Studie könnte angeführt werden, dass die hier verwendeten Kriterien (und damit Ergebnisse) lediglich repräsentativ für schulische Bewertungsprozesse sind. Diese Tatsache stellt jedoch prinzipiell kein Problem dar, da es durchaus denkbar wäre, dass verschiedene

Expertengruppen aufgrund unterschiedlicher Zielsetzungen und Arbeitsschwerpunkte (z.B. Kinder zum Schreiben von Texten auf einem Mindestmaß an Alltagstauglichkeit trainieren, die Qualität von Texten zur Veröffentlichung bewerten, die sprachliche Präzision eines Textes beurteilen) unterschiedliche Kriterien (und eventuell andere als die hier angeführten) verwenden (Eckes, 2008; Sakyi, 2000) oder diese unterschiedlich gewichten, aber innerhalb ihrer Gruppe dennoch zu ähnlichen Bewertungen der Relevanz der Kriterien und auch der Textqualität kommen. Tatsächlich zeigt die vorliegende Studie jedoch, dass innerhalb aller Gruppen die Varianz der Bewertungen hoch und damit die Konsistenz der Bewertungen gering ist. Keine der untersuchten Gruppen gibt zuverlässige Textbewertungen ab; die Bewertenden innerhalb einer Gruppe unterscheiden sich hinsichtlich ihrer Strengemaße stark voneinander. Darüber hinaus herrscht in keiner der Gruppen Einigkeit darüber, wie die vorgegebenen Kriterien bei der Textbewertung zu gewichten sind.

Entlang dieser Befunde ergibt sich damit die Problematik, dass – entsprechend der vorab definierten Kriterien – keine der untersuchten Gruppen als Expertengruppe im Bereich der Textbewertung betrachtet werden kann. Dementsprechend muss einerseits festgestellt werden, dass Lehrkräfte nicht über das notwendige Maß an diagnostischer Expertise verfügen, welches sie zur Förderung der Textproduktionskompetenz ihrer Schüler*innen benötigen; andererseits, auch das zeigen die Befunde, stellt die Bewertung der Textqualität jedoch nicht nur für Lehrkräfte eine besondere Herausforderung dar. Auch Wissenschaftler*innen und Lektor*innen und Autor*innen fehlt die notwendige Expertise bei der Bewertung der Textqualität. Sie können damit nicht als gute Orientierungspunkte für Lehrkräfte gelten. Entlang der präsentierten Befunde scheint es keine sinnvollen Anker/gute Vorbilder zu geben, an denen sich Lehrkräfte bei der Textbewertung orientieren könnten.

Dennoch sollte aus diesen Ergebnissen nicht der Schluss gezogen werden, dass eine objektive und zuverlässige Bewertung der Textproduktionskompetenz grundsätzlich unmöglich ist und die Verwendung von Textbewertungskriterien generell zu unzuverlässigen Urteilen führt. Vielmehr zeigen die Ergebnisse auf, dass im Kontext der Textproduktionskompetenz (und deren Bewertung) noch ein großer Bedarf an Grundlagenforschung besteht.

Vorrangiges Ziel weiterer Forschungsbemühungen sollte die Identifikation und Definition manifester Minimalanforderungen (!) für einen guten Text sein, die Lehrkräfte im Kontext schulischer Bewertungsprozesse (diagnostischer Fragestellungen) nutzen können. Zur Identifikation solcher Minimalstandards sollte zum einen die bisherige empirische Befundlage berücksichtigt, darüber hinaus aber auch die Expertise von Angehörigen verschiedener inhaltlicher

Disziplinen mit Bezug zur Textproduktion (z.B. Wissenschaft, Journalismus, Didaktik) einbezogen werden. Aspekte von Textqualität/Textproduktionskompetenz, hinsichtlich deren Relevanz Experten sich einig sind, wären ein erster Ansatzpunkt für konkrete Forschungsvorhaben und Operationalisierungsversuche. Einige solcher Kriterien lassen sich entlang der vorliegenden Daten – anhand der Übereinstimmungen zwischen den verschiedenen Expertengruppen – bereits identifizieren. Eine weitere Identifikationsstrategie zur Definition von Minimalstandards stellt die Orientierung an Faktoren dar, die die Konsistenz und Verständlichkeit eines Textes stören und damit seine Funktionstüchtigkeit beeinträchtigen. Ausgehend von der Annahme, dass ein Text in erster Linie funktional sein sollte, lassen sich so eventuell minimale Gelingensbedingungen identifizieren, welche nicht das Höchstmaß an Textproduktionskompetenzen erfordern und damit vor allem bei der Diagnostik akuter Schreibschwierigkeiten relevant sein könnten. Diese Idee stellt gerade vor dem Hintergrund bisheriger Forschungsansätze, welche Textproduktionskompetenzen und Schreibprozesse zumeist auf Basis von Schreibexperten modellieren (siehe Kapitel 6), einen innovativen neuen Zugang dar. Eine wichtige Frage, die es bei der Suche nach solchen Minimalstandards zu klären gilt, ist, inwiefern diese textsortenspezifisch sind. Denkbar ist, dass aufgrund der unterschiedlichen Anforderungen, welche Textmuster an Schreiber*innen stellen (z.B. Argumentation vs. Erzählung – siehe dazu Kapitel 7.1.1.2) auch die Minimalanforderungen zwischen Texten unterschiedlicher Textarten variieren. Zur Klärung dieser Frage und der Weiterentwicklung der aufgeworfenen Ideen braucht es entsprechende empirische Studien.

5.7 Limitationen der vorliegenden Studie

Bei der Interpretation und Diskussion der vorliegenden Daten und Ergebnisse gilt es einige Einschränkungen zu beachten. Eine Einschränkung der Daten besteht darin, dass sich die Gewichtungen der Bewertungskriterien nur auf die vordefinierten Kriterien als solche und nicht auf einen konkreten Bewertungsprozess, d.h. deren konkrete Anwendung, beziehen. Die Daten lassen folglich nur Rückschlüsse darauf zu, welches Gewicht die Bewertenden den Kriterien im Allgemeinen geben. Es besteht jedoch die Möglichkeit, dass die tatsächlich vorgenommenen Gewichtungen der Kriterien bei der Bewertung eines Textes von den in dieser Studie angegebenen Gewichtungen abweichen. Weitere Studien im Bereich der Textbewertung sollten sich daher auf die tatsächliche Anwendung der Bewertungskriterien beziehen und deren Gewichtung/Relevanz für das Zustandekommen eines Gesamturteils zur Textqualität untersuchen. Forschungsansätze hierzu werden im Ausblick dieser Arbeit (Kapitel 8) thematisiert.

Eine weitere Einschränkung der Studie ergibt sich daraus, dass die untersuchten Kriterien in Anlehnung an existierende (zumeist theoretische) Literatur zu den Themen Textqualität, Schreibkompetenz und Textbewertung definiert wurden. Es ist jedoch möglich, dass die Wissenschaft zum jetzigen Zeitpunkt noch nicht alle für die Beurteilung der Textqualität relevanten Kriterien identifizieren konnte und die hier vorgegebenen Kriterien das Konstrukt nicht vollständig abbilden. Darüber hinaus führt die Orientierung an bisheriger Forschungsliteratur und bestehenden Bewertungsrastern dazu, dass auch die hier vorgegebenen Kriterien nicht eindeutig operationalisiert, d.h. nicht manifest und eindeutig zu bewerten sind. Verschiedene Bewerter*innen haben diese daher möglicherweise unterschiedlich interpretiert, was wiederum eine mögliche Erklärung für die hohe Varianz innerhalb der einzelnen (Experten-)Gruppen sein kann.

Außerdem lassen sich einige der vorgegeben Kriterien (z.B. Kriterium 10 „Unterstützung der Lesenden durch den Aufbau des Textes“ und Kriterium 11 „Berücksichtigung der Perspektive der Lesenden“) theoretisch als „übergeordnete“ Kriterien klassifizieren (siehe Kapitel 7.1.1), die sich hinsichtlich verschiedener anderer Aspekte (hier Kriterien) auf die Gestaltung eines Textes auswirken können. Empirische Studien dazu, wie hoch die Relevanz der hier vorgegebenen Kriterien im Einzelnen tatsächlich ist und ob sich eine Hierarchisierung dieser ergibt, sind dringend erforderlich, um die vorliegenden Befunde besser einordnen zu können.

5.8 Zwischenfazit

Die Befunde der in diesem Kapitel beschriebenen Studie zeigen: Die Bewertung der Textqualität stellt nicht nur für Lehrkräfte eine besondere Herausforderung dar, auch Wissenschaftler*innen und Lektor*innen und Autor*innen fehlt es an Expertise bei der Bewertung. Keine der untersuchten Gruppen ist dazu in der Lage, zuverlässige Urteile zur Textqualität abzugeben. Ziel weiterer Forschungsbemühungen muss daher sein, das Konstrukt der Textproduktionskompetenz einer psychometrischen Kriterien entsprechenden Messung zugänglich zu machen. Nur wenn dies gelingt, kann die (schulische) Diagnostik und Förderung der Textproduktionskompetenz verbessert werden.

Voraussetzung für eine objektive, reliable und valide Messung eines Konstruktes ist immer dessen Definition und konkrete Operationalisierung. Hierzu soll die vorliegende Arbeit einen Beitrag leisten. Bevor jedoch erste Versuche und Möglichkeiten zur Definition und Operationalisierung der Textproduktionskompetenzen vorgestellt werden, wird im folgenden Kapitel (Kapitel 6) der aktuelle Stand der nationalen und internationalen Schreibforschung skizziert,

welcher eine wichtige Grundlage für die vorliegende Arbeit und die in Kapitel 7 dargelegte Modellierung der Textproduktionskompetenz sowie die daran anschließende Studie (Studie 2 dieser Arbeit) darstellt.

6. Modellvorstellung zur (gelungenen) Textproduktion

Im Laufe der Zeit haben sich in der Forschung verschiedene Modellvorstellungen dazu entwickelt, wie ein (gelungener) Schreibprozess abläuft, welche Anforderungen die Textproduktion an die Schreibenden stellt und wie sich die erforderlichen Kompetenzen zur Textproduktion entwickeln. Entlang ihrer Ausrichtung lassen sich die existierenden Modelle in Prozessmodelle, kognitive Modelle und Entwicklungsmodelle unterteilen. Aus jeder der genannten Kategorien wird im Folgenden je ein Modell vorgestellt und diskutiert.

6.1 Das Schreibprozess Modell von Hayes und Flower

Das wohl bekannteste Modell zum Schreibprozess stammt aus dem Jahr 1980 und wurde von den beiden Wissenschaftler*innen John Hayes und Linda Flower entwickelt. Auf Basis von Daten, die mithilfe der Technik des lauten Denkens bei kompetenten Schreiber*innen gewonnen wurden, entwickelten sie ein Modell, welches aus drei Komponenten besteht: 1) der Schreibumgebung, 2) dem Langzeitgedächtnis (LZG) und 3) dem eigentlichen Schreibprozess.

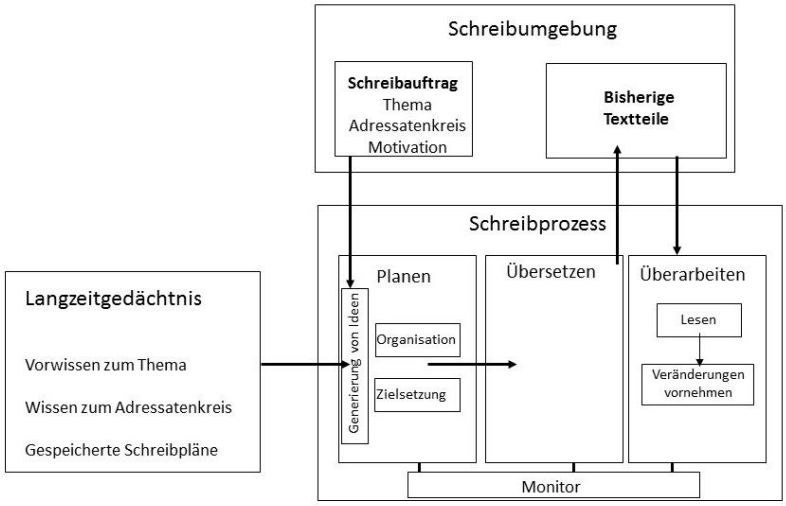


Abbildung 7: Das Schreibprozessmodell (eigene Darstellung; in Anlehnung an Hayes & Flower, 1980)

Bei 1) der Schreibumgebung handelt es sich um den Kontext, in den der jeweilige Schreibprozess eingebettet ist. Die Schreibumgebung beinhaltet den Schreibanlass (z.B. in Form einer Schreibaufgabe), welcher wiederum Angaben zum thematischen Kontext und dem Kreis der Adressanten beinhaltet. Außerdem werden in der Schreibumgebung die bereits im Verlauf des Schreibprozesses produzierten Textteile verortet, wodurch sich die Schreibumgebung während des Schreibprozesses fortlaufend verändern kann. Auf die Informationen innerhalb der Schreibumgebung wird während des Schreibprozesses kontinuierlich zugegriffen. Gleiches gilt für 2) das Langzeitgedächtnis, denn hierin sind potentielle Informationen zu verschiedensten Themenfeldern, zum Adressatenkreis und – bei erfahrenen Schreiber*innen – auch Schreibpläne bzw. Textmuster sowie zugehörige standardisierte Formulierungen gespeichert.

Die Bedeutung, welche der Schreibumgebung und dem Langzeitgedächtnis im 3) Schreibprozess zukommt, wird bei dessen genauerer Betrachtung klar. Der Schreibprozess selbst ist in drei Subprozesse – das Planen, das Übersetzen und das Überarbeiten – unterteilt. Diese drei Subprozesse werden von einem kognitiven Monitor koordiniert.

Inhalt der **Planungsphase** sind drei eng miteinander verzahnte Elemente: a) die Generierung von Ideen, b) die Organisation/Strukturierung und Hierarchisierung dieser Ideen und c) die Entwicklung einer Zielsetzung. Zur *Generierung von Ideen* werden Informationen zu einem bestimmten Themenfeld aus dem Langzeitgedächtnis abgerufen und genutzt. Je besser der entsprechende Themenbereich im Langzeitgedächtnis gefüllt und organisiert ist, desto leichter erfolgt der Zugriff hierauf während des Schreibens. Die auf diese Weise generierten Ideen werden im Prozesselement der *Ideenorganisation* entsprechend ihrer Relevanz ausgewählt/beibehalten oder verworfen und anschließend hierarchisiert bzw. sortiert. Auch hierbei spielt das Langzeitgedächtnis eine Rolle, da es Informationen zum Leserkreis liefert, die für die Auswahl und Hierarchisierung der Ideen/Inhalte relevant sein können (Wer benötigt welche Informationen?). Parallel zu den Vorgängen der Ideengenerierung und -organisation erfolgt die *Entwicklung der Zielsetzung*. In dieser wird festgehalten, welche Funktion der (intendierte) Text erfüllen soll. Zur Festlegung der Zielsetzung müssen Informationen aus der Schreibumgebung (Schreibanlass, Themenfeld, Adressatenkreis etc.) berücksichtigt werden. Die entwickelte Zielsetzung wirkt sich auf die Prozesselemente der *Ideengenerierung und Ideenorganisation* aus, denn hieran ausgerichtet werden bestimmte Informationen aus dem Langzeitgedächtnis abgerufen (oder nicht abgerufen) und entwickelte Ideen/Inhalte weiterverfolgt oder eben (vorerst) verworfen.

Gleichzeitig können aber auch durch die Entwicklung von Ideen und deren Organisation Veränderungen in der Zielsetzung vorgenommen werden. Die Elemente der Planungsphase interagieren demnach miteinander, eine klare Abfolge dieser lässt sich nicht bestimmen und auch im Rahmen verschiedener Schreibprozesse der gleichen Autorin/des gleichen Autors können diese Prozesse in unterschiedlichem Maße interagieren (z.B. bedarf es beim Verschriftlichen eines Kochrezeptes, welches die Autorin/der Autor schon mehrfach selbst angewandt hat, weniger Planungsarbeit als beim Verfassen einer schriftlichen Abschlussarbeit zu einem komplexen theoretischen Sachverhalt). Als Ergebnis der Planungsphase entstehen (gedankliche oder in schriftlichen Plänen festgehaltene) Netzwerke von Ideen und Plänen für den intendierten Text. Planungsphasen können im Schreibprozess an verschiedenen Stellen auftreten und sind nicht, wie die Anordnung der Modellelemente möglicherweise nahelegt, auf den Beginn des Schreibprozesses begrenzt.

Die im Rahmen der Planungsphase entstandenen Pläne bzw. Ideennetzwerke müssen in der **Übersetzungsphase** in einen konventionell schriftlichen Text überführt werden. Hierzu werden Gedanken (Ideennetze) in geschriebene Sprache übersetzt und verschriftlicht. Adäquate schriftsprachliche Formulierungen müssen gefunden werden, um die erdachten Ideen/Inhalte in Form eines funktionstüchtigen Textes bzw. funktionstüchtiger Textteile zu präsentieren. Hierbei kommt dem Langzeitgedächtnis erneut eine wichtige Funktion zu, denn der Übersetzungsprozess kann entlastet werden, indem beispielsweise bereits gespeicherte Schreibpläne aus dem Langzeitgedächtnis abgerufen und/oder standardisierte (und gespeicherte) Formulierungsmuster aus dem Langzeitgedächtnis zur Übersetzung genutzt werden können. Ist der Zugriff auf das Langzeitgedächtnis erschwert, zum Beispiel durch die fehlende Automatisierung von Transkriptionsfähigkeiten (Handschrift/Orthografie) und einer damit verbundenen hohen Belastung des Arbeitsgedächtnisses, kann dies den Übersetzungsprozess wiederum stark beeinträchtigen. Voraussetzung für gelungene Schreibprozesse sind demnach möglichst große Kapazitäten des Langzeitgedächtnisses (bzw. Arbeitsgedächtnisses). Basale Prozesse, wie beispielsweise das Transkribieren oder Lesen, sollten demnach möglichst automatisiert ablaufen.

An die Phase der Übersetzung – oder aber auch der Planung – knüpft die **Überarbeitungsphase** an. In dieser Phase werden bereits produzierte Textteile – oder auch der ganze Text – von der/dem Schreiber*in gelesen und mit der bestehenden Zielsetzung verglichen. Stimmen Zielsetzung bzw. Planung und der vorliegende Text nicht miteinander überein, wird der Text (im Idealfall) entsprechend der Zielsetzung editiert/verändert. Überarbeitungsprozesse müssen sich jedoch nicht zwangsläufig auf bereits verschriftlichte Textteile beziehen. Es ist ebenso

möglich, dass noch nicht verschriftlichte Ideen verändert und/oder Pläne entlang bereits verschriftlichter Elemente verändert werden. Überarbeitungsprozesse können demnach Veränderungen des bisherigen Planes, d.h. weitere Überarbeitungen notwendig machen. Sie können daher auch als „Sprungbrett“ für weitere Planungs- und Übersetzungsprozesse betrachtet werden.

Diese Tatsache verdeutlicht eine Grundannahme des Modells: Die Phasen des Schreibprozesses laufen parallel zueinander ab und können zu jeder Zeit im Textproduktionsprozess aktiv werden. Beispielsweise die Überarbeitungsphase ist demnach nicht auf den Abschluss des Schreiprozesses begrenzt, vielmehr werden im Rahmen von neuen Planungs- und Übersetzungsprozessen auch immer wieder neue Überarbeitungsprozesse in Gang gesetzt, welche wiederum neue Planungs- oder Übersetzungsphasen nach sich ziehen können.

Die Aktivitäten der jeweiligen Subprozesse und deren Zusammenspiel werden vom sogenannten *Monitor* koordiniert. Wie diese Koordination abläuft, d.h. was bei der Textproduktion im Kopf der schreibenden Person zur Koordination der verschiedenen Teilprozesse geschieht, bleibt jedoch hinter dem Begriff des Monitors verborgen. Hierbei handelt es sich um eine Schwäche des Prozessmodells, denn es bleibt sehr nah am Textentstehungsprozess und bildet in Teilprozessen ab, was bei der Textproduktion auf dem Papier passiert. Kognitive Vorgänge und die dazu notwendigen Ressourcen bzw. konkreten Textproduktionskompetenzen werden nicht modelliert.

6.2 Das kognitive Schreibmodell von Hayes

Mit dem kognitiven Schreibmodell aus dem Jahr 2012 begegnet Hayes dieser Kritik. Er greift in diesem Modell die Fragestellung auf, was im Kopf der schreibenden Person passiert. Das Modell ist weniger fokussiert auf den Ablauf des Schreibprozesses; die/der Schreiber*in und deren/dessen Kognition stehen im Mittelpunkt (Jakobs & Perrin, 2014). Schreiben wird hier nun als ein Zusammenspiel verschiedener (kognitiver) Handlungen verstanden, welche es zielführend aufeinander abzustimmen gilt. Die für das (erfolgreiche) Schreiben notwendigen Teilelemente werden zueinander in Beziehung gesetzt und im Modell auf drei verschiedenen Ebenen verortet – der Ressourcen-, der Prozess- und der Kontrollebene.

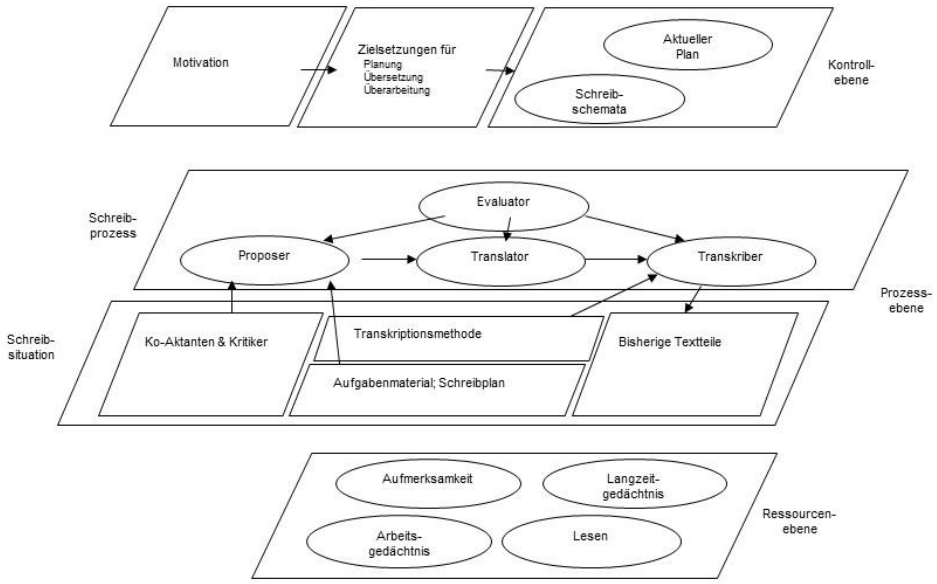


Abbildung 8: Das kognitive Schreibmodell (eigene Abbildung nach Hayes, 2012)

Die unterste Ebene des Modells bildet die **Ressourcenebene**. Auf dieser Ebene sind (kognitive) Fähigkeiten verortet, welche wichtige Voraussetzungen für das Schreiben darstellen. Zu diesen Fähigkeiten zählen: die Aufmerksamkeit, das Arbeitsgedächtnis, die Lesekompetenz und das Langzeitgedächtnis. Die Ressourcenebene bildet die Basis des Modells und damit die Basis für eine erfolgreiche Bewältigung des Problemlöseprozesses beim Schreiben von Texten. Die hier verorteten Elemente können als Voraussetzung für gelungene Textproduktionsprozesse betrachtet werden.

Darüber liegt, in der Mitte des Modells, die **Prozessebene**. Diese setzt sich aus der Schreibumgebung und dem Schreibprozess zusammen. Die Schreibumgebung entspricht in weiten Teilen der Idee des Modells von 1980. Teil dieser sind die Transkriptionstechnologie, die Schreibaufgabe und zugehöriges Aufgabenmaterial, der potentielle Adressatenkreis und bereits geschriebene Textteile (hierzu zählen auch: schriftlich festgehaltene Schreibpläne, z.B. in Form von Gliederungen). Bei den meisten der auf dieser Ebene beschriebenen Elemente handelt es sich um Vorgaben oder äußere Bedingungen, mit denen die schreibende Person umgehen muss und die es beim Schreiben zu berücksichtigen gilt. Allerdings wird in diesem Modell dem potentiellen Adressatenkreis (hier als „Ko-Akteuren und Kritikern“ benannt) die Möglichkeit eingeräumt, sich (kritisch) zum Text zu äußern (z.B. können Lehrpersonen Feedback an Schü-

ler*innen geben), um so Überarbeitungsprozesse anzuregen. Neben bereits geschriebenen Textteilen nimmt in diesem Modell somit ein weiteres Element der Schreibumgebung direkten Einfluss auf den Schreibprozess.

Der Schreibprozess, welcher ebenfalls auf der Prozessebene verortet ist, wird im vorliegenden Modell nicht in die drei Subprozesse Planen, Übersetzen und Überarbeiten aufgeteilt. Stattdessen werden auf der Ebene des Schreibprozesses kognitive Elemente benannt, durch deren Nutzung der Schreibprozess gesteuert und ausgeführt werden kann. So gibt es zum Beispiel einen *Evaluator*, der dafür zuständig ist, zu ermessen, was dem Schreibprozess bzw. dem Ziel dienlich ist und was nicht. Diese Überprüfung erfolgt aber nicht zwangsläufig in einer Überarbeitungsphase, sondern fortlaufend; die vom *Proposer* generierten Ideen, vom „Translator" übersetzten Gedanken und die vom *Transkriber* verschriftlichten Textelemente werden im gesamten Verlauf des Schreibprozesses immer wieder evaluiert. Alle Teilelemente des Schreibprozesses sind über den *Evaluator*, aber auch untereinander verbunden. Weitere Verbindungen bestehen zu Elementen der Schreibumgebung. So ist beispielsweise der *Proposer* eng mit der Schreibaufgabe und dem Adressatenkreis verbunden, sodass (im Idealfall) Ideen vom *Proposer* generiert werden, die eng hierauf abgestimmt sind.

Koordiniert bzw. kontrolliert werden die Prozess- und die Ressourcenebene von der Kontrollebene. Auf dieser finden sich die Motivation, die Zielsetzung, aktuelle (gedachte) Schreibpläne und Textmusterkenntnisse (gespeicherte Schreibpläne und/oder standardisierte Formulierungen). Allen hier verorteten Elementen ist gemein, dass sie eine leitende Funktion beim Schreiben haben. So leitet beispielsweise die Zielsetzung den Schreibprozess: Bereits geschriebene Textteile werden im Hinblick auf das Ziel ständig vom *Evaluator* geprüft. Dieser greift hierzu immer auf den aktuell bestehenden (aber veränderlichen) Text und den aktuell bestehenden (aber veränderlichen) Schreibplan zurück und nutzt außerdem beispielsweise Vorwissen zu bestimmten Textmustern, um die bisherigen Formulierungen (durch den *Translator*) im vorliegenden Text zu bewerten. Grundlegende Voraussetzung zur Durchführung dieser Prozesse bilden die benannten kognitiven Fähigkeiten (z.B. das Lesen und die Aufmerksamkeit) auf der Ressourcenebene.

Im Modell angelegt sind die Grundvoraussetzungen gelungener Textproduktionsprozesse und die damit verbundene Idee, dass kompetente Schreiber*innen immer wieder neu aus einem Set unterschiedlicher Handlungsoptionen deren zieldienlichste Kombination auswählen, um den jeweiligen Schreibprozess bestmöglich zu bewältigen. Gute Schreiber*innen haben demnach die Kompetenz, die unterschiedlichen Teilbereiche des Schreibens immer neu/anders,

aber möglichst zielführend miteinander zu kombinieren. Ein gelungener Schreibprozess bzw. ein guter Text ist damit am Ende ein gelungenes Zusammenspiel der benannten Elemente auf Kontroll-, Prozess- und Ressourcenebene.

Wie genau dieses Zusammenspiel (im Inneren der schreibenden Person) angelegt sein muss, damit es zielführend ist, lässt sich anhand des Modells jedoch nicht eindeutig bestimmen, da verschiedene Schreibprozesse immer abhängig von unterschiedlichsten Faktorkombinationen (z.B. persönliche Schreibstile, Vorwissen, Aufgabenstellungen, Textsorten etc.) und somit nicht einheitlich beschreibbar sind. Diese Tatsache stellt jedoch laut dem Autor keine Schwäche des Modells dar, sondern spiegelt die Komplexität des Schreibens wider.

Aus der diagnostischen Perspektive handelt es sich hierbei allerdings um eine gravierende Schwäche des Modells, denn aufgrund dessen hoher Flexibilität lassen sich die Bedingungen eines gelungenen Zusammenspiels nicht eindeutig benennen. Und auch auftretende Schwierigkeiten bzw. deren Ursachen können nicht eindeutig und differenziert bestimmt werden. Dies hat zur Folge, dass es an Anhaltspunkten für Unterricht und Förderung fehlt. Denn entlang des Modells ist ein „schlechter“ Text kein Indikator für allgemein mangelnde Schreibkompetenz, sondern erst einmal nur ein Hinweis darauf, dass bei der Ausführung einer der Teilhandlungen innerhalb des komplexen Prozesses der Textproduktion etwas nicht gelungen ist. Was genau nicht gelungen ist, worin die/der Schreiber*in also noch Unterstützung bzw. Übung braucht, bleibt offen. Notwendige Voraussetzung für eine differenzierte Diagnostik von vorhandenen bzw. (noch) nicht vorhandenen Schreibkompetenzen wäre die konkrete Benennung von Teilkompetenzen, welche Schreiber*innen zur Produktion funktionstüchtiger Texte beherrschen müssen. Die Benennung von konkreten schreibspezifischen Teilkompetenzen erfolgt im vorliegenden Modell nicht; der Fokus liegt hier eher auf den kognitiven Anforderungen und Prozessen während der Textproduktion.

6.3 Schreibentwicklungsmodelle

Entwicklungsmodelle bieten wichtige Anhaltspunkte zur Diagnostik von Textproduktionskompetenz, da hier die Entwicklung von (Schreib-)Kompetenzen entlang der Norm (d.h. für den Großteil der Population der Schreiber*innen bzw. der Lerner*innen, die keine Abweichungen im negativen Sinne aufweisen) dargestellt wird. Im Kontext der Schreibkompetenz sind diese Modelle so angelegt, dass sie die Entwicklung beschreiben, welche Schreibnovizen

durchlaufen, bis sie als kompetente Schreiber*innen dazu in der Lage sind, den Problemlöseprozess des Schreibens (angelehnt an Hayes & Flower, 1980) erfolgreich bewältigen zu können.

Das wohl bekannteste Schreibentwicklungsmodell stammt von Bereiter (1980). Es modelliert die Entwicklung der Schreibkompetenz über fünf Stufen. Zu Beginn dieser Entwicklung (Stufe 1 „associative writing“, S. 83) verschriftlichen Schreibende Ideen bzw. Inhalte rein assoziativ und verbinden diese, wenn überhaupt, nur geringfügig miteinander. Grund für diese Vorgehensweise ist die Tatsache, dass in dieser Phase der Prozess des Verschriftlichens (Handschrift und Rechtschreibung) noch viele kognitive Kapazitäten benötigt und somit im Vordergrund steht. Erst im zweiten Entwicklungsschritt (Stufe 2 „performative writing“, S. 85*)* rückt dann der Text selbst bzw. das Textprodukt mehr in den Fokus. Texte werden nun entsprechend schriftsprachlicher Konventionen und Normen verfasst. Syntaktische Strukturen der geschriebenen Sprache, (erste) Elemente von Textmustern, aber auch orthografische Elemente wie die Interpunktion, werden von Schreiber*innen erkannt bzw. erlernt und beim Schreiben zunehmend berücksichtigt. In der darauffolgenden Stufe (Stufe 3 „communicative writing“, S. 86) nimmt der/die Schreiber*in zusätzlich die Lesenden in den Blick. Eine entscheidende Komponente der Schreibkompetenz, die Fähigkeit zur Adressatenorientierung, entwickelt sich. Besonderheiten der schriftlichen Kommunikation (*zerdehnte Kommunikation*; Ehlich, 1984) und die damit einhergehenden Bedürfnisse der Lesenden sowie die Bedürfnisse bestimmter Adressatenkreise (z.B. vorhandenes/nicht vorhandenes Vorwissen) werden von der/dem Schreiber*in antizipiert und fließen so in die Gestaltung des Textes mit ein. Daran anschließend bzw. eng mit dieser Fähigkeit verknüpft entwickelt sich die Fähigkeit, den eigenen Text kritisch zu betrachten (Stufe 4 „unified writing“, S. 87). Die/der Schreiber*in setzt sich kritisch mit dem produzierten Text oder einzelnen Passagen darin auseinander und überprüft diese hinsichtlich ihrer Funktionalität. Durch einen Abgleich mit der Intention der Schreiberin/des Schreibers bzw. deren/dessen Zielsetzung können so Textstellen identifiziert werden, die einer Überarbeitung bedürfen. Diese Überarbeitungen werden mit dem Ziel vorgenommen, den Text möglichst funktionstüchtig zu gestalten. Auf der letzten Entwicklungsstufe (Stufe 5 „epistemic writing“, S. 87) entdecken Schreiber*innen dann die lernförderliche Funktion des Schreibens: Lerngegenstände können über eine Verschriftlichung erfasst und miteinander in Verbindung gesetzt werden. Denkprozesse können durch die Verschriftlichung unterstützt und vorangetrieben werden, der Erkenntnisgewinn wird somit gefördert. Schreiber*innen, die diese Stufe der Entwicklung erreicht haben und

das Schreiben von Texten als Werkzeug einsetzen können, werden als „Schreibexperten“ bezeichnet.

Etwas einfacher, aber inhaltlich vergleichbar, beschreiben Bereiter und Scardamalia (1987) den entwicklungsbezogenen Einsatz von Schreibstrategien. Unterschieden werden hier nicht fünf aufeinander aufbauende Entwicklungsstufen, sondern die Verwendung von zwei Schreibstrategien: zum einen die des „knowledge telling“ (S.5) und zum anderen die des „knowledge transforming“ (S.6). Schreibende, die die *knowledge-telling*-Strategie verfolgen (meist Novizen), geben ihre Ideen ungeordnet wieder und planen die Struktur des Textes nicht. Ideen sind mitunter nicht aufeinander abgestimmt, da Schreibende vorangegangene und nachfolgende Informationen nicht berücksichtigen bzw. zueinander in Beziehung setzen. Der Fokus bei der Textproduktion liegt im Rahmen des *knowledge telling* vor allem auf der Generierung von Ideen und damit Inhalten für den Text. Dies stellt in diesem Stadium der Entwicklung eine besondere Herausforderung dar, weil Schreibende (im Gegensatz zu mündlicher Kommunikation) ohne eine(n) aktive(n) Kommunikationspartner*in selbstständig Ideen und Inhalte generieren müssen. Hierzu orientieren sie sich zunächst am Erlebten und geben dies wieder. Schreiber*innen, die die Strategie des *knowledge transforming* verfolgen, bewältigen den Problemlöseprozess des Schreibens hingegen erfolgreich und verfassen einen funktional angemessenen Text. Im Gegenteil zu Schreibnovizen nutzen sie sprachliche Mittel, um den produzierten Text sowohl auf lokaler als auch auf globaler Ebene kohärent und damit verständlich für die Adressaten zu gestalten. Informationen werden gezielt ausgewählt, d.h. auf die Zielsetzung des Textes und die Bedürfnisse der Lesenden abgestimmt. Diese Strategie des *knowledge transforming* wird mit zunehmender Schreiberfahrung und -expertise immer häufiger angewandt, auch wenn im Prozess der Schreibentwicklung noch Mischformen der Strategienutzung existieren.

Die Entwicklung von Expertise im Bereich der Schreibkompetenz reicht bis in die Phase der Adoleszenz und kann sich auch hier noch fortsetzen (z.B. durch das Hinzukommen neuer Anforderungen hinsichtlich bisher unbekannter Textmuster). Ein Ende der Entwicklung lässt sich demnach nicht genau bestimmen und auch die Definition starrer Altersgrenzen, zu denen bestimmte Entwicklungsphasen „normalerweise“ abgeschlossen sein sollten, ist nicht möglich, da Unterricht und Förderung starken Einfluss hierauf nehmen (Becker-Mrotzek & Böttcher, 2012). Orientierungswerte bietet jedoch die von Becker-Mrotzek & Böttcher (2012) vorgenommene Unterteilung der Schreibentwicklung in vier Phasen. Entlang dieser Phasen nehmen Schreiber*innen im Alter zwischen fünf und sieben Jahren erste Schreibversuche vor,

die jedoch noch stark von basalen Prozessen dominiert sind. Im Alter von sieben bis zehn Jahren bauen Schreiber*innen diese Fähigkeiten aus und beginnen mit dem Schreiben erster längerer Texte. Hierbei verfolgen sie die Strategie des *knowledge telling*. Erst im Alter von zehn bis 14 Jahren beginnen Schreiber*innen, sich Gedanken um die Bedürfnisse der Lesenden zu machen und verwenden (zumindest in Teilen) die Strategie des *knowledge transforming*. Schreiber*innen bringen nun ihr Sachwissen, grammatisches Wissen und Wissen zu Textarten in den Schreibprozess ein. Die Phase der literalen Orientierung, d.h. der Schreibexpertise, erreichen Schreiber*innen dann mit der Adoleszenz.

Schreibentwicklungsmodelle lassen demnach eine (ungefähre) Einordnung dazu zu, welche Kompetenzen Schreiber*innen in einem bestimmten Entwicklungsalter erworben haben (sollten). Im Rahmen diagnostischer Prozesse können Entwicklungsmodelle somit Anhaltspunkte für mögliche Entwicklungsverzögerungen liefern. Diese Möglichkeit muss – vor dem Hintergrund, dass im deutschen Sprachraum derzeit kein psychometrisches Testverfahren zur Einordnung von Schreibkompetenzen anhand von Normdaten zur Verfügung steht – durchaus gewürdigt werden. Kritisch zu betrachten bleibt jedoch, dass auch im Rahmen von Entwicklungsmodellen keine manifesten Merkmale des Konstruktes der „Schreibkompetenz“ bestimmt werden und es demnach schwierig zu beurteilen bleibt, inwiefern Schreiber*innen über die auf den Entwicklungsstufen verorteten Strategien verfügen oder nicht.

6.4 Möglichkeiten und Grenzen bisheriger Modelle

Zusammenfassend lässt sich sagen, dass bisherige Modelle, trotz der angeführten Kritik, wichtige Erkenntnisse zum Schreibprozess selbst, den kognitiven Prozessen und Anforderungen des Schreibens sowie der Entwicklung der Textproduktionskompetenz bereitstellen. Die auf Basis der Modellvorstellungen gewonnenen Erkenntnisse haben darüber hinaus sowohl in der Vergangenheit als auch in der Gegenwart weitere gewinnbringende Forschungsarbeiten und die fortlaufende Weiterbearbeitung und Neuentwicklung von Modellen angeregt. Auch die in Kapitel 2 aufgeführten Definitionen orientieren sich an den vorgestellten Modellen.

Kritisch muss jedoch angemerkt werden, dass den existierenden Modellen (sowie den in Kapitel 2 dieser Arbeit aufgeführten Definitionen zur Schreibkompetenz) gemein ist, dass diese keine eindeutige Operationalisierung des Konstruktes der Textproduktionskompetenz

zulassen und damit die eindeutige und zuverlässige Messung dieser Kompetenz aus methodischer Perspektive nicht gewährleistet werden kann (siehe hierzu auch Kapitel 4). Die Entwicklung eines empirisch abgesicherten, theoretisch fundierten und analytisch widerspruchsfreien Kompetenzmodells, welches es zur Diagnostik und Förderung der Schreibkompetenz braucht, steht zum aktuellen Zeitpunkt (2020) damit noch aus (siehe auch Ossner, 2006; Becker-Mrotzek & Schindler, 2008).

Unter Berücksichtigung der Befunde der in Kapitel 5 vorgestellten Studie erscheint die Konzeption eines solchen Kompetenzmodells, auf dessen Basis eine zuverlässige Erfassung der Textproduktionskompetenz möglich wird, jedoch dringend erforderlich zu sein. Im Rahmen der zweiten Studie dieser Dissertation wird der Versuch unternommen, ein solches Modell zu konzipieren. Aufbauend auf diesem Modell wird ein neuartiges Verfahren vorgestellt, welches die Erfassung der Textproduktionskompetenz entlang psychometrischer Kriterien ermöglichen soll.

7. Schreibkompetenz standardisiert erfassen (Studie 2)

In diesem Kapitel wird ein neues Verfahren[4] zur Messung der Textproduktionskompetenz vorgestellt. Präsentiert werden grundlegende Ansätze zur Entwicklung eines (förderdiagnostischen) Instrumentes, das den in Kapitel 3 genannten Bedarfen zur Diagnostik und Förderung der Textproduktionskompetenz gerecht werden soll.

Das vorgestellte Verfahren bezieht sich schwerpunktmäßig auf informierende Texte, da diese entsprechend den Lehrplänen in der adressierten Altersgruppe (4.-9. Jahrgangsstufe) bekannte Textsorten beinhalten, die über die Spanne der Jahrgangsstufen hinweg durchgängig relevant sind (Becker-Mrotzek & Böttcher, 2012). Diese Altersgruppe wurde gewählt, da nach den Modellen zur Entwicklung der Schreibkompetenz ab dem 4. Schuljahr davon auszugehen ist, dass basale Fähigkeiten grundlegend erworben sind und Schreiber*innen die kommunikative Funktion der Textproduktion funktional nutzen können (Bereiter, 1980; Bereiter & Scardamalia, 1987). Zusätzlich stellt der Wechsel von der Grundschule in die weiterführenden Schulen einen wichtigen Übergang in der schulischen Entwicklung dar. In weiterführenden Schulen werden schriftliche Prüfungsformate zunehmend relevant und die Schreibkompetenz spielt in

[4] Das hier vorgestellte Verfahren wird derzeit von der Autorin gemeinsam mit Kolleg*innen verschiedener Disziplinen entwickelt. Hierzu zählen: Prof. Dr. Michael Becker-Mrozek, Prof. Dr. Jörg Jost, Prof. Dr. Markus Linnemann, Prof. Dr. Christian Rietz, Prof. Dr. Alfred Schabmann und Dr. Barbara Schmidt. Weitere Informationen siehe Vorwort.

verschiedenen Schulfächern eine entscheidende Rolle. Zudem stellt das Ende der Sekundarstufe I für einen Teil der Schüler*innen das Ende der schulischen Laufbahn dar. Speziell im sonderpädagogischen Feld könnte eine Erhebung der Textproduktionskompetenz Hinweise für eine nötige (zusätzliche) Förderung geben.

7.1 Das Konvergenzmodell

Wichtige Voraussetzung für die zuverlässige Erfassung einer Kompetenz ist ein Basismodell, welches die Grundlage für die konkrete Operationalisierung des zu erfassenden Konstruktes bildet. Hierzu wurde, orientiert an den Modellvorstellungen von Hayes und Flower (1980) sowie Hayes (2002) und den darauf aufbauenden Forschungsbefunden, das *Konvergenzmodell* (KM; Hennes et al., 2018) entwickelt. Neu am KM ist die Tatsache, dass einzelne konkrete Dimensionen der Textproduktionskompetenz postuliert werden („Textproduktionskompetenz im engeren Sinne"; Hennes et al., 2018, S. 298), auf die Schreibende zurückgreifen, um einen funktionstüchtigen Text zu produzieren. Diese Teilkompetenzen sind fester Bestandteil des KM und ihr Einfluss auf den Textproduktionsprozess wird modelliert. Das – bisher auf Basis abstrakter Prozessvariablen definierte – Konstrukt der Textproduktionskompetenz soll auf diese Weise operationalisiert werden.

Der Fokus des KM liegt, wie der Name sagt, auf der Tatsache, dass die Textproduktion hier als Prozess der Über-/Bearbeitung von vorläufigen Textteilen oder Schreibideen hin zu einem „optimalen" Text verstanden wird. Der Schreibprozess wird (nach dem Modell von Hayes & Flower, 1980) als eine iterative Abfolge von Planung und Übersetzung (das eigentliche Schreiben) betrachtet. „Triebfeder" dieses Prozesses ist die ständige Über-/Bearbeitung des Plans und/oder der Übersetzung auf Basis des (noch nicht perfekten) bisherigen Textes. Planung, Übersetzung und Über-/Bearbeitung werden von drei kognitiven Instanzen auf einer Metaebene kontrolliert und koordiniert. Die Benennung und die Funktionsweisen dieser Instanzen wurden aus dem Modell von Hayes (2012) übernommen.

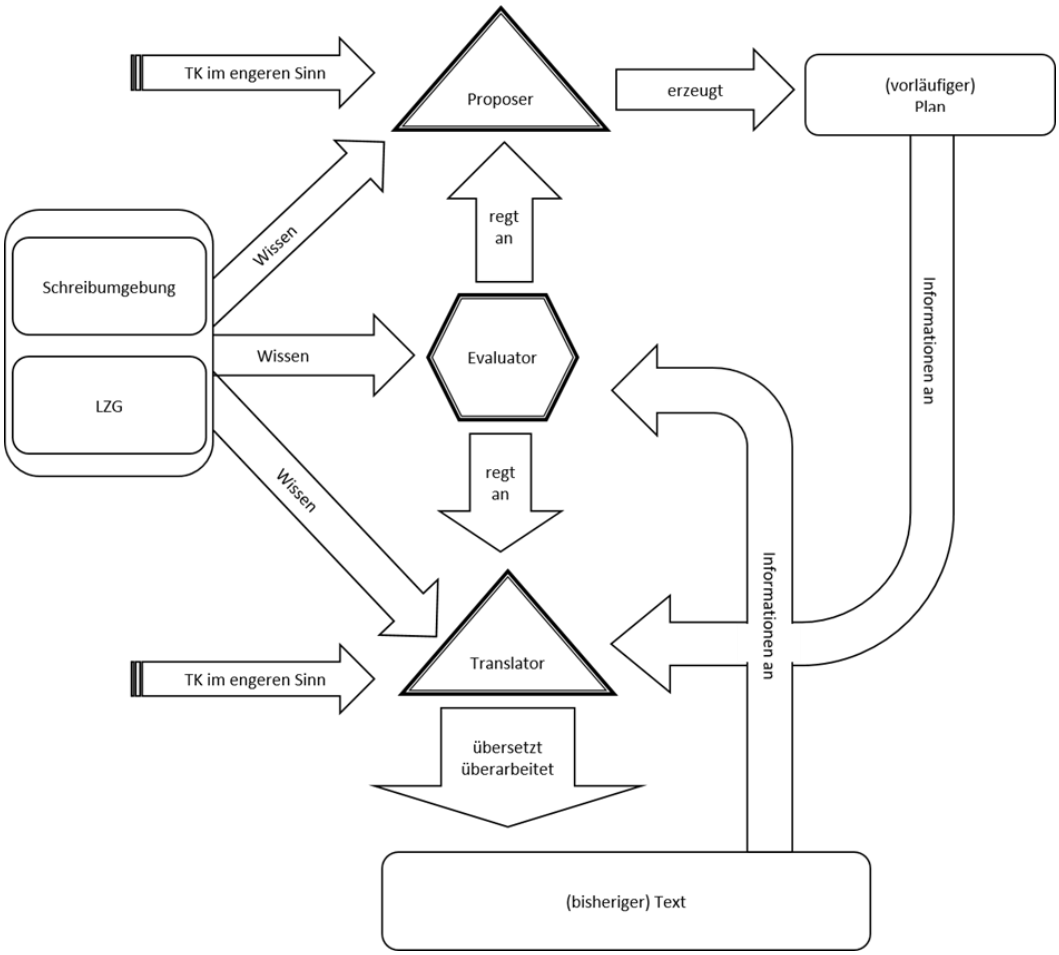


Abbildung 9: Das Konvergenzmodell (theoretisches Basismodell der Testkonstruktion)

Die oberste Instanz im KM ist der *Evaluator*, der zwei ihm untergeordnete Instanzen (*Proposer* und *Translator*) überwacht und diese hinsichtlich der Gesamtzielsetzung des Textes steuert. Der *Proposer* ist die primär kreative Instanz im Schreibprozess. Er liefert einen im Idealfall auf Ziel und Adressat abgestimmten „Vorschlag“ für den Inhalt und die Grobstruktur des Textes (als vorläufiger Schreibplan). Hierzu berücksichtigt er Informationen (z.B. über den Rezipienten) aus dem Langzeitgedächtnis der Schreiberin/des Schreibers und aus der Schreibumgebung (Hayes & Flower, 1980). Der entstandene Schreibplan kann in dieser ersten Arbeitsschleife durchaus noch unscharf („fuzzy“; Dam-Jensen & Heine, 2013, S. 93) sein (z.B. erste unklare Ideen, Halbsätze etc.). Beim Schreiben kann dann immer wieder auf den *Proposer* zurückgegriffen werden, was eine ständige Konkretisierung des vorläufigen Schreibplans nach sich zieht. Der *Translator* steuert den Übersetzungsprozess. Er übersetzt die Elemente des Plans in konventionelle (Schrift-)Sprache. So entsteht ein (vorläufiger) Text, der sichtbares Ergebnis der Planung und konkrete Basis für (eine erste oder weitere) Überarbeitungsschleife(n) ist, welche vom *Evaluator* angeregt werden. Dieser überprüft die vorläufigen Schreibprodukte und

initiiert gegebenenfalls die Überarbeitung des Plans oder aber des Textes. Überarbeitungen, d.h. vorgenommene Veränderungen, stehen in Wechselwirkung: Veränderte Textteile können neue Planungsprozesse nach sich ziehen und umgekehrt (Hayes & Flower, 1980; MacArthur & Graham, 2016). Abgeschlossen ist der Iterationsprozess dann, wenn die/der Schreiber*in keine Veränderungsnotwendigkeit mehr sieht (wobei sich die/der Schreiber*in auch mit einer vorläufigen Fassung zufriedengeben kann).

7.1.1 Textproduktionskompetenz im engeren Sinn

Die Umwandlung von Ideen in Text erfolgt auf Basis einzelner (kleinteiliger) Arbeitsschritte. Ergänzend zu den (gerade beschriebenen) Komponenten des Schreibprozessmodells und des kognitiven Modells finden sich im KM daher zusätzlich konkrete Kompetenzen, auf die Schreiber*innen zurückgreifen, um einen funktionstüchtigen Text zu produzieren. Diese Kompetenzen werden als *Textproduktionskompetenz im engeren Sinn* (TPK) bezeichnet. Auf der Ebene der TPK erhebt das Modell (noch) keinen Anspruch auf Vollständigkeit. Auf hierarchieniedrige Komponenten (z.B. Transkription und Orthografie) wird in der vorliegenden Definition bewusst verzichtet, weil sie im Rahmen des KM (nach Sturm & Weder, 2016) als Voraussetzungen für die Textproduktionskompetenz (TPK) im engeren Sinne betrachtet werden. Die TPK setzt sich aus den Dimensionen der Adressatenorientierung, des Textmusterwissens, der Informationslogistik, der Nutzung kohärenzstiftender Mittel sowie der sprachlichen Kreativität zusammen, die im Folgenden näher beschrieben werden.

7.1.1.1 Die Adressatenorientierung

Texte, die ihre kommunikative Funktion über Raum und Zeit hinweg erfüllen sollen (*zerdehnte Kommunikation*; Ehlich, 1983), müssen von Schreiber*innen so angelegt sein, dass Leser*innen auf Basis der schriftlich vorliegenden Informationen (und entsprechendem Vorwissen) ein *mentales Modell* („mental model“, Johnson-Laird, 1980, S. 73; „situation model“ van Dijk & Kintsch, 1983, S. 290; van Dijk, 1995, S. 394) zu den Inhalten des Textes aufbauen können. Hierzu müssen Schreiber*innen die Bedürfnisse von Leser*innen antizipieren. Erst entlang solcher Antizipationsleistungen und deren Ergebnissen können Schreibende Texte dann entsprechend der Bedürfnisse potentieller Leser*innen gestalten.

Notwendige Voraussetzung für die Adressatenorientierung ist demnach zunächst die Fähigkeit zur Perspektivübernahme (Becker-Mrotzek et al., 2014; Schmitt, 2011). Diese Fähigkeit

ist im Laufe des gesamten Textproduktionsprozesses relevant, denn Schreiber*innen müssen hier immer wieder die Perspektive von Leser*innen auf den (bisherigen) Text und dessen inhaltliche wie sprachliche Merkmale einnehmen können. Sie müssen die Interpretation des Textes durch die Lesenden fortlaufend modellieren (Kellogg, 2008) und aufbauend hierauf Bearbeitungs- oder auch Überarbeitungsschritte vornehmen.

Die Fähigkeit zur fortlaufenden Berücksichtigung des modellierten Adressaten („imagined reader“ Kellogg, 2008, S. 5) entwickelt sich jedoch erst mit zunehmender Schreibexpertise. In einem von Kellogg (2008) postulierten dreistufigen Entwicklungsmodell sind der Phase der Adressatenorientierung zwei weitere Entwicklungsphasen vorangestellt. In der ersten dieser beiden Phasen („knowledge telling“, S. 4; siehe auch: Bereiter & Scardamalia, 1987) liegt der Fokus der Textproduktion auf den Absichten der Autorin/des Autors und deren/dessen Ideen, Plänen und Zielen. In der zweiten dieser Phasen („knowledge transforming“, S. 4; siehe auch: Bereiter & Scardamalia, 1987) entwickelt sich zusätzlich zu den eigenen Absichten (Model 1) eine mentale Modellvorstellung zum Text aus Sicht der Autorin/des Autors (Model 2). Erst in der dritten Entwicklungsphase („knowledge crafting“, S. 4) bildet sich dann eine stabile Modellvorstellung zur Leserin/zum Leser aus (Model 3; siehe auch Flower, 1994). Erst hier entwickeln Schreibende die Fähigkeit zur fortlaufenden Antizipation der Leser*innen, deren Sicht auf den Text und deren Bedürfnisse und erst jetzt können Schreibende ihre Absichten (Modell 1) und bisher produzierte Textteile (Modell 2) auch mit der antizipierten Sicht der Leser*innen hierauf (Modell 3) in Verbindung bringen und eigene Ziele sowie die Bedürfnisse von Leser*innen gleichzeitig berücksichtigen und dementsprechend adressatengerechte Texte formulieren.

Die Fähigkeit zur Antizipation der Adressaten (Perspektivübernahme) – eine in erster Linie kognitive Leistung (Kellogg, 2008; Schindler, 2004) – bildet die wichtigste Voraussetzung der Adressatenorientierung. Abstrakte Denkprozesse (Piaget, 1973) sowie ausreichende Kapazitäten des Arbeitsgedächtnisses (Kellogg, 2008) sind zur Realisierung der Adressatenorientierung erforderlich. Die erforderlichen Kapazitäten des Arbeitsgedächtnisses sind laut Kellog (2008) auch der Grund dafür, dass sich eine stabile Repräsentation der Leser*innenperspektive erst in einer späten Phase der Schreibentwicklung ausbilden kann. Dennoch finden, wenn die Kapazitäten des Arbeitsgedächtnisses dies zulassen, auch in früheren Entwicklungsphasen gelegentlich Antizipationsprozesse statt und die Bedürfnisse von Adressaten werden in bestimmten Teilen des Textproduktionsprozesses berücksichtigt.

Konkret realisiert wird die Adressatenorientierung anhand sprachlicher Mittel auf Textebene. Zur Formulierung adressatengerechter Texte stehen Schreiber*innen verschiedene

sprachliche Mittel zur Verfügung, die sie entsprechend der Bedürfnisse des Adressaten nutzen. So variieren Schreibende zum Beispiel die Informationsdichte eines Textes, d.h. den Anteil explizit gegebener Informationen, entsprechend des Vorwissens der adressierten Leser*innen, nutzen Konnektoren zur Verdeutlichung von Relationen oder strukturieren Texte entsprechend bekannter Muster. Zur Realisierung der Adressatenorientierung auf Textebene bedarf es somit der übrigen Komponenten der TPK im engeren Sinne (siehe unten). Der Adressatenorientierung kommt daher im Konvergenzmodell eine übergeordnete Funktion im Hinblick auf die weiteren Komponenten der TPK zu (siehe dazu auch Becker-Mrotzek & Schindler, 2007). Wie genau die Adressatenorientierung Einfluss auf die übrigen Komponenten nimmt, wird in den folgenden Erläuterungen derselben deutlich.

7.1.1.2 Textmusterwissen

Textmuster stellen „vorgeformte[] textuelle[] Strukturen“ (Luginbühl & Perrin, 2011, S. 7) dar, die für eine Gruppe bestimmter Texte (z.B. Erzähltexte) gelten (siehe auch Heinemann, 2000). Sie liefern Anhaltspunkte zur prototypischen Organisation und Struktur einer bestimmten Textsorte, aber auch zu möglichen Inhalten, deren Abfolge und der zu realisierenden kommunikativen Funktion eines Textes. Es handelt sich demnach um Modelle, die es bei der Textproduktion sprachlich zu füllen gilt (Bachmann & Becker-Mrotzek, 2017; Feilke, 2010). Hierzu existieren bestimmte prototypische, konstitutive Textprozeduren (zum Beispiel Textbausteine und prototypische Formulierungen; Bachmann & Becker-Mrotzek, 2017; Pohl, 2007).

Textmuster können, samt zugehöriger Prozeduren, als Schemata (Bachmann & Becker-Mrotzek, 2017) betrachtet werden, an denen sich Schreibende bei der Produktion eines Textes orientieren und entlang derer sie den Text und seine Bestandteile organisieren und strukturieren können (Richard & Schmidt, 2002). Beispielsweise können Ideen und Informationen anhand eines bereits bekannten Musters und dessen notwendiger Bestandteile zunächst gezielter generiert und dann entsprechend strukturiert werden (Hennes, Büyüknarci, Rietz & Grünke, 2015). Die Orientierung an einem solchen Schema und die Verwendung entsprechender Prozeduren erleichtern das Schreiben eines Textes, denn Schreibende können sich bei der Übersetzung ihrer Ideen(netze) vornehmlich auf die inhaltliche Gestaltung des Textes konzentrieren (Becker-Mrotzek & Schindler, 2007).

Allerdings unterscheiden sich die Anforderungen, welche verschiedene Textmuster an die Schreibenden stellen, stark voneinander. So erfordern beispielsweise Erzählungen nur we-

nig Veränderungen oder Umstrukturierungen am eigenen Wissen, denn das zugehörige Textmuster orientiert sich am unmittelbaren „Erlebniswissen“ (Becker-Mrotzek & Schindler, 2007, S. 16). Ideen können hier recht einfach entlang der tatsächlichen Ereignisse wiedergegeben werden. Andere Textmuster (beispielsweise Argumentationen) erfordern eine starke Umstrukturierung des bereits vorhandenen Wissens; verschiedene Wissensinhalte müssen entlang der Textfunktion möglichst zielführend zu einem logischen und schlüssigen Informations-/Argumentationsstrang kombiniert werden (Becker-Mrotzek & Schindler, 2007). Das Anforderungsniveau ist hier entsprechend höher. Dennoch entlastet das Wissen um Textmuster und deren prozedurale Elemente die/den Schreibende(n) im Schreibprozess (Hayes & Flower, 1980; Hayes, 2012).

Das Wissen um Textmuster und deren Prozeduren liegt im Langzeitgedächtnis. Schreibende können diese entsprechend des Schreibanlasses bzw. der Schreibaufgabe, wenn diese ausreichend gefestigt sind, abrufen und realisieren (Becker-Mrotzek & Schindler, 2007). Doch auch wenn prototypische Textmuster im Langzeitgedächtnis gespeichert und somit verinnerlicht sind, können diese nicht einfach abgerufen und in jedem Schreibprozess gleichermaßen routiniert angewandt werden. Bei der Anwendung von Textmustern muss der „Gebrauchszusammenhang“ (Feilke, 2010, S. 3) berücksichtigt werden, d.h. bestehende Textmuster müssen entlang des jeweiligen Schreibanlasses bzw. Schreibprozesses adaptiert werden. Hierzu sind kreative Veränderungen und damit Abweichungen vom musterhaften Vorgehen erforderlich (Luginbühl & Perrin, 2011). Wichtigste Voraussetzung für eine solche Adaption ist ein tiefes Verständnis von Textmustern, denn nur auf dessen Basis können Muster gebrochen und zum eigenen Zweck verändert werden. Hierbei müssen Schreibende jedoch die Bedürfnisse der Adressaten beachten, denn neben der leitenden Funktion beim Schreiben erfüllt ein Textmuster auch für Lesende eine leitende Funktion (Pohl, 2007): Diese können entlang der ihnen bekannten Textstruktur Gelesenes organisieren und erinnern. Kenntnis über Textmuster und entsprechende Strukturierungen des Textes erleichtern das Leseverständnis (Shanahan et al., 2010). Lesende nutzen ihr Wissen zum Textmuster für den Aufbau ihres mentalen Modells (Kintsch, 1988; Strong, 2019). Die von Schreibenden vorgenommenen Adaptionen eines Textmusters sollten demnach nur so weit reichen, wie es den Lesenden nicht irritiert oder gar bei der Konstruktion des mentalen Modells behindert (Adressatenorientierung) – es sei denn, genau dies wäre vom Schreibenden intendiert.

7.1.1.3 Informationslogistik

Bei der Textproduktion stehen Schreiber*innen vor der Herausforderung, den Inhalt des Textes sinnvoll und für den Adressaten verständlich zu gestalten. Während des gesamten Schreibprozesses müssen Schreibende hierzu entlang des Textthemas und ihrer Zielsetzung mögliche Inhalte für den Text in Form von Ideen bzw. Informationen generieren. Aus den generierten Informationen müssen dann die für die Lesenden relevanten Informationen ausgewählt und weniger relevante Informationen verworfen werden (Bereiter, 1980; Bereiter & Scardamalia, 1987; Hayes & Flower, 1980; Hayes, 2012). Gelingt dies, können Schreibende – entsprechend der Grice'schen Maximen (1975)[5] für kommunikative Beiträge – Texte so informativ wie nötig, gleichzeitig nicht überinformierend und damit relevant und angemessen für Lesende gestalten.

Aber nicht nur die Informationsdichte ist relevant für die Produktion eines adressatengerechten Textes, auch die Organisation der gewählten Informationen trägt maßgeblich zur Verständlichkeit eines Textes bei. Aufeinander aufbauende Informationen sollten in entsprechender Reihenfolge gegeben werden (geordnet; Grice, 1975), Widersprüche vermieden (Grice, 1975) und Anforderungen des Textmusters berücksichtigt werden (z.B. die enge Orientierung am schrittweisen Vorgehen im Falle einer Instruktion; Becker-Mrotzek, 2003). Gelingt dies, werden die Erwartungen der Lesenden (an den Text bzw. die/den Schreibende*n) erfüllt und deren mentale Modellbildung unterstützt. Zusätzlich ermöglicht eine gute Informationsorganisation, dass Lesende sich Informationen – auch wenn diese nicht explizit im Text gegeben werden – logisch erschließen können (Implikaturen; Grice, 1975; Harabagiu & Moldovan, 2000; Linke & Nussbaumer, 2000). Hierzu nutzen Lesende ihr Vorwissen zum Thema (Holle, 2010) und setzen voraus, dass Schreibende sich bei der Produktion eines Textes auf den „Normalfall" („default"; Clark, Hodges, Stephan & Moldovan, 2005, S. 1) beziehen und aktiv an das Weltwissen der Lesenden („common ground"; Krifka & Musan, 2012, S. 6) anknüpfen. Weichen Schreibende von diesem Normalfall ab, so gilt es, Informationen explizit im Text zu geben und solche Abweichungen sprachlich entsprechend zu markieren.

[5] Aktuellere Studien (Sripada, Reiter, Hunter & Yu, 2003) zur empirische Überprüfung der Funktionalität von Texten, die künstlich, d.h. anhand einer Computersoftware generiert wurden, zeigen, dass die Grice'schen Maximen immer noch wichtige Prinzipien für die gelungene Informationsweitergabe in Textform darstellen. Leser, hier sowohl User als auch Experten, bezogen sich bei ihren Verbesserungsvorschlägen im Hinblick auf die Informationslogistik der Texte ausschließlich auf Aspekte der von Grice (1975) postulierten Maximen.

Die Informationsweitergabe in Textform ist damit ein kooperativer Prozess: Schreibende gestalten Texte so, dass Lesende bei der mentalen Modellbildung möglichst optimal unterstützt werden, dürfen dabei aber eine gewisse Fähigkeit der Lesenden dazu voraussetzen, sich bestimmte (implizite) Informationen eigenständig erschließen zu können – natürlich immer unter Berücksichtigung des entsprechenden Vorwissens beim Adressatenkreis.

Die Fähigkeit zur Perspektivübernahme bildet demnach auch für die Informationslogistik eine wichtige Voraussetzung, denn nur auf Basis der Antizipation des Adressatenkreises und dessen Vorwissen können Schreibende eine Entscheidung über die Relevanz von Informationen für Lesende treffen und den eigenen Text auf die notwendige Informationsdichte und die Logik der präsentierten Informationen hin überprüfen.

7.1.1.4 Die Nutzung kohärenzstiftender Mittel

Der Begriff der Textkohärenz bezeichnet „die semantisch-konzeptuelle Kontinuität“ (Schwarz-Friesel, 2006, S. 64) eines Textes, oder einfacher ausgedrückt, dessen logischen, inhaltlichen Zusammenhang (Averintseva-Klisch, 2013). Hierbei kann zwischen *globaler Kohärenz* auf Ebene des Gesamttextes und *lokaler Kohärenz* auf Satzebene unterschieden werden. Damit ein Text seine kommunikative Funktion erfüllen kann, sollten Schreiber*innen einen Text produzieren, der auf beiden Ebenen kohärent und damit für die Lesenden nachvollziehbar ist. Hierzu stehen Schreibenden verschiedene sprachliche Mittel zur Verfügung, anhand derer sie Zusammenhänge zwischen Informationen/Ideen auf der Textoberfläche explizit sichtbar machen können. Der hierdurch erzeugte explizite sprachliche Zusammenhang auf der Textoberfläche wird als Kohäsion bezeichnet, während sich der Begriff der Kohärenz vornehmlich auf die mentalen Konstruktionsleistungen von Lesenden bezieht, welche diese auf Grundlage des Textes vornehmen (Averintseva-Klisch, 2013; Bachmann, 2002; McNamara, Kintsch, Songer, & Kintsch, 1996; O’Reilly & McNamara, 2007; Schwarz-Friesel, 2006).

Sprachliche Mittel zur Herstellung von Kohäsion werden als Kohäsionsmittel bezeichnet (Crossley & McNamara, 2010). Entlang ihrer Funktion lassen sich Kohäsionsmittel in zwei Gruppen unterteilen. Die erste Gruppe bilden Kohäsionsmittel zur Herstellung lokaler Kohäsion, hierunter fallen: Konnektoren, d.h. Konjunktionen, Adverbien und Partikel im engeren Sinne (z.B. Becker-Mrotzek et al., 2014, Crossley & McNamara, 2010); semantische Relationen lexikalischer Art, z.B. Meronyme und Synonyme oder die Verwendung von Wörtern eines semantischen Feldes; Rekurrenzen, d.h. die Wiederverwendung von Wörtern, Phrasen oder Konzepten (z.B. Crossley & McNamara, 2010); Referenzen, d.h. Pro-Formen aller Art (z.B.

Averintseva-Klisch, 2013) sowie Anaphern (z.B. Becker-Mrotzek et al., 2014; Crossley & McNamara, 2010). Die zweite Gruppe bilden sprachliche Mittel zur Herstellung globaler Kohäsion. Hierzu zählen textstrukturierende Mittel (z.B. Überschriften), aber auch die kontinuierliche Berücksichtigung und Fortführung inhaltlicher Aspekte (z.B. Textthema und Makropropositionen) dienen dem Gesamtzusammenhang des Textes und unterstützen die Lesenden dabei, Kohärenz mental herzustellen.

Textkohärenz muss jedoch nicht immer explizit durch Kohäsionsmittel hergestellt werden. Es ist möglich, dass einzelne Sätze (oder auch ganze Texte) für die Lesenden kohärent sind, ohne dass Kohäsionsmittel verwendet werden (Becker-Mrotzek et al., 2014; Rickheit & Strohner, 2003). Gleichzeitig können einzelne Sätze kohäsiv sein, ohne dass der Text als Gesamtprodukt dieser Sätze kohärent für die Lesenden ist (Schwarz, 2001).

Ob ein Text kohärent für die Lesenden ist oder nicht hängt jedoch nicht nur vom Text selbst und den darin angelegten Pfaden und expliziten Verknüpfungen ab. Auch die Art und Weise, wie eine Person einen Text liest und interpretiert, wirkt sich entscheidend auf die Herstellung von (mentaler) Kohärenz aus (Todd et al., 2004). Studien zeigen, dass das Vorwissen und die Lesefähigkeiten einen enormen Einfluss auf die mentale Herstellung von Kohärenz nehmen (McNamara et al., 1996; O'Reilly & McNamara, 2007). So profitieren beispielsweise schwache Leser*innen von dem Einsatz von Kohäsionsmitteln (McNamara, Louwerse, McCarthy & Graesser, 2010), während dies für gute Leser*innen, deren Vorwissen umfangreich ist, nicht unbedingt der Fall sein muss, da diese die Fähigkeit haben, Zusammenhänge herstellen zu können, ohne explizit auf diese hingewiesen zu werden. Der Einsatz von Kohäsionsmitteln kann bei guten Leser*innen sogar den Prozess der mentalen Kohärenzbildung stören (O'Reilly & McNamara, 2007).

Das Vorliegen von Textkohärenz beziehungsweise die darauf aufbauende Herstellung mentaler Kohärenz ist damit etwas sehr Subjektives (Todd et al., 2004). Einer Messung zugänglich sind lediglich die Merkmale auf Textoberfläche, d.h. die verwendeten (oder nicht verwendeten) Kohäsionsmittel.

7.1.1.5 Sprachliche Kreativität

Der Prozess des Schreibens wird häufig auch als *Problemlöseprozess* (Flower & Hayes, 1980) beschrieben. Schreiber*innen benötigen zur Bewältigung dieses Problemlöseprozesses die Fähigkeit, Lösungen für auftretende Probleme im Textproduktionsprozess zu finden, die

über die standardmäßige Verwendung von schriftsprachlichen Mittel (z.B. typische Formulierungen) hinausgehen. Hierzu bedarf es neben Fleiß und Übung kognitiver sowie sprachlich kreativer Fähigkeiten (divergentes Denken; Dem-Jensen & Heine, 2013; Kellogg, 2008).

Bei der sprachlichen Kreativität handelt es sich um eine bisher kaum erforschte Teildimension der Textproduktionskompetenz. Eine Definition dieser Komponente kann daher nicht vorgenommen werden. In Anlehnung an die Kreativitätsforschung lassen sich jedoch Dimensionen des Konstruktes der Kreativität benennen und auf die Textproduktion übertragen. Charakteristisch für kreatives Handeln ist die Fähigkeit dazu, etwas Neuartiges beziehungsweise *Originelles* und gleichzeitig *Zweckmäßiges* zu erarbeiten (Amabile, 1982; Sternberg, 1999; Sternberg, Kaufman & Pretz, 2002). Sprachliche Kreativität im Kontext der Textproduktion zeigt sich somit darin, dass Schreiber*innen etwas Neuartiges, d.h. zunächst Unerwartetes und damit Originelles verschriftlichen (Jost & Böttcher, 2012). Hierbei kann es sich beispielsweise um neuartige Inhalte („inhaltliche Wagnisse"; Böttcher & Becker-Mrotzek, 2003, S. 54), originelle Formulierungen/Darstellungen („neue sprachliche Wege"; Nussbaumer & Sieber, 1995, S. 41; „außergewöhnliche […] sprachliche Mittel"; Böttcher & Becker-Mrotzek, 2003, S. 54) sowie unerwartete strukturelle bzw. formelle Vorgehensweisen („formale Wagnisse"; Nussbaumer & Sieber, 1995, S. 41) handeln. Genauer spezifizieren lassen sich die benannten Aspekte sprachlicher Kreativität aufgrund ihrer Eigenschaft, neuartig beziehungsweise unerwartet zu sein, jedoch a priori nicht. Möglich ist lediglich eine Annäherung an sie anhand von Beispielen. So können formelle/strukturelle Wagnisse beispielsweise kreative Adaptionen des Textmusters sein (Luginbühl & Perrin, 2011) und „neue sprachliche Wege" (Nussbaumer & Sieber, 1995, S. 41) über Vergleiche führen. Allen Wagnissen ist jedoch gemein, dass ihr Einsatz – unter der Prämisse, dass die/der Schreibende einen funktionstüchtigen Text produzieren möchte – immer unter Berücksichtigung der Bedürfnisse der Lesenden erfolgen sollte. Nur so kann sprachliche Kreativität zweckmäßig sein.

Das Erkennen und Bewerten solcher kreativer Momente erfordert ein hohes Maß an Expertise im Bereich der Textproduktion, denn eindeutige Bewertungskriterien hierzu können nicht formuliert werden. Bewertende müssen daher auf ihre eigene Erfahrung und ihre impliziten Kenntnisse zu sprachlicher Kreativität zurückgreifen (siehe Kapitel 5.1).

7.2 Umsetzung im Test

7.2.1 Grundprinzip der Testkonstruktion

Aus dem KM wird deutlich, dass Textproduktionskompetenzen auf verschiedenen Ebenen erfasst werden können, nämlich auf der Ebene der Schreibkompetenz im Sinne von Hayes und Flower (1980) bzw. Hayes (2012) sowie auf der Ebene der im Modell ausgewiesenen TPK im engeren Sinn. Im angestrebten Verfahren steht der Zugang über die TPK im engeren Sinne im Fokus. Dieser Ansatz erscheint deshalb geeignet, weil über die TPK im engeren Sinne ein klarer Rahmen für die Diagnostik und Förderung gegeben ist. Stärken und Schwächen von Schreiber*innen lassen sich differenziert identifizieren (d.h. es kann getrennt erfasst werden, inwiefern ein/e Schreiber*in bereits über Textmusterwissen verfügt und dieses anwenden kann und inwiefern ihr/ihm der Einsatz von Kohäsionsmitteln gelingt). Basierend auf diesen differenzierten Befunden können dann konkrete Interventionen durchgeführt werden, die beispielsweise besonders auf eine Erweiterung des Textmusterwissens ausgerichtet sind, wenn in diesem Bereich Schwächen identifiziert wurden. Über den Zugang der TPK können förderdiagnostische Fragestellungen somit präzise beantwortet und Interventionen konkret auf identifizierte Förderbedarfe abgestimmt werden. Eine Diagnostik auf Basis der TPK im engeren Sinne hat somit aus förderdiagnostischer Perspektive deutliche Vorteile gegenüber der Messung der Textproduktionskompetenz anhand der Textqualität, die genau dies – wie einleitend dargelegt – nicht ermöglicht.

7.2.1 Ein konkretes Aufgabenbeispiel

Im Folgenden wird beispielhaft an einer Skala zur Informationslogistik (adäquate Informationsdichte) dargestellt, wie die verschiedenen Einzelaufgaben zur Erfassung der unterschiedlichen Facetten/Teildimensionen der TPK grundsätzlich konstruiert sind. Die hier vorgestellte Skala soll messen, inwiefern die Testpersonen (TP) in der Lage sind, vollständige Informationen zu geben. Hierbei besteht die Aufgabe darin, für unwissende Adressat*innen eine Bauanleitung zu einem Gebäude aus handelsüblichen Bauklötzen zu schreiben (siehe Abbildung 10). Den TP liegen hierzu eine Abbildung des Bausteingebäudes in Form eines Fotos sowie ein Set der für das jeweilige Gebäude verwendeten Bausteine vor.

Folgende Instruktion wird gegeben:

„Du siehst auf der Abbildung ein Gebäude aus Bausteinen.

1. Schau dir das Bild genau an. Baue das Gebäude nach.

2. Überprüfe, ob du alles richtig gebaut hast. Achte besonders auf diese Stelle: (Markierung der explizit zu beschreibenden Stelle im Gebäude; siehe unten)

3. Ein Freund soll das Gebäude nachbauen. Er hat das Bild nicht. Schreibe ihm eine genaue Anleitung dazu, wie er das Gebäude auf dem Bild nachbauen kann!“

Für diese Aufgabe wird ein Set von fünf Bausteingebäuden mit unterschiedlicher Komplexität konstruiert. Die Konstruktion der Gebäude erfolgte anhand folgender Prinzipien:

• Jedes Gebäude besteht aus Steinen, die hinsichtlich Farbe und Form nur gemeinsam variieren. So besteht für die Schreibenden die Möglichkeit, die Steine anhand ihrer Farbe zu benennen. Geometrische Bezeichnungen müssen nicht bekannt sein. Die Relevanz des (mathematischen) Vorwissens wird so reduziert.

• Die Komplexität der Bausteingebäude nimmt mit jeder Schwierigkeitsstufe zu. Pro Stufe wird ein Baustein mehr verbaut. Dieser zusätzliche Stein erforderte beim Schreiben eine oder mehrere neue Instruktionen in Bezug auf dessen Lage und Position. Mit zunehmender Komplexität erhöht sich demnach die adäquate Informationsdichte.

• Daten aus Pilotversuchen in verschiedenen Jahrgangsstufen (4-9) zeigten, dass die TP sehr viel implizites Wissen des Adressaten voraussetzten und nur sehr grobe Informationen verschriftlichten. Aus diesem Grund enthält in der vorliegenden Version jedes Gebäude mindestens ein besonderes Element (kritisches Element), dessen Position oder Ausrichtung im Text detaillierter als bei anderen Elementen beschrieben werden muss (z.B. der hochkant gestellte Stein in Abbildung 10).

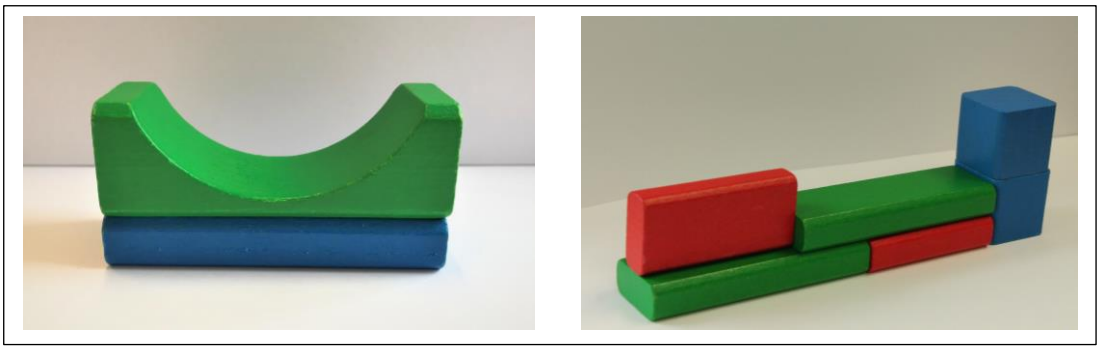


Abbildung 10: Beispiel eines Gebäudes der niedrigsten (links) und der höchsten Komplexitätsstufe (Originalabbildungen). Beide Gebäude enthalten explizit zu instruierende Elemente. Im linken Gebäude muss der grüne Stein (häufig als „Brücke“ bezeichnet) als „mit den Füßen nach oben“ instruiert werden; im rechten Gebäude muss der obere rote Stein „auf die lange, schmale Seite gestellt“ werden.

Zwei Aspekte waren bei der Konstruktion der Aufgabe wesentlich. Erstens sollten Unterschiede im Vorwissen der Testpersonen (TP) ausgeglichen werden. Dies war nötig, da entsprechend dem KM der *Proposer* auf Vorwissen aus dem Langzeitgedächtnis (LZG) des Schreibers/der Schreiberin zugreift und so das Vorwissen die Messung der Schreibkompetenz beeinflusst. Das sachliche Vorwissen wurde daher a priori im Test vorgegeben, indem den TP die für das zu instruierende Gebäude notwendigen Bausteine vorlagen und diese dazu aufgefordert waren, das Gebäude in einem ersten Schritt selbst zu bauen. Außerdem wurde die Aufgabe so konstruiert, dass der Adressatenkreis des zu verfassenden Textes eindeutig benannt und das Wissen der adressierten Person explizit beschrieben wurde. Dieser Schritt war notwendig, da die Informationen, die Schreibende geben müssen, stark vom Adressatenkreis des Textes abhängen.

In Kapitel 7.4 werden erste Ergebnisse zur internen Konsistenz der Skala, zur Konstruktvalidität sowie zur Entwicklung der erreichten Testergebnisse über verschiedene Klassenstufen berichtet. Es bestehen folgende Erwartungen:

1) Interne Konsistenz und faktorielle Validität: Angesichts der Kürze der Skalen wird ein (zufriedenstellendes) Cronbachs α von ~ 0,80 erwartet. Die ansteigende Komplexität der Bausteingebäude sollte sich im Testscore monoton abbilden. Zudem wird erwartet, dass sich die Aufgaben durch eine gemeinsame latente Dimension („Informationsdichte“) erklären lassen.

2) Konvergente und divergente Validität: Wichtig war es, die Aufgaben so zu konstruieren, dass die Raumvorstellung der TP bei der Bearbeitung keine Rolle spielt. Es wurde darauf geachtet,

dass die Rechts-Links-Orientierung der Steine irrelevant war. Zudem waren alle Steine während der Bearbeitung der Schreibaufgabe verfügbar. Entsprechend wird eine Nullkorrelation zwischen dem Testscore der Aufgabe und der Raumvorstellung erwartet. Hingegen sind Korrelationen von Schreibaufgaben mit dem basalen Lesen (Ahmed, Wagner & Lopez, 2014), dem Wortschatz (Castillo & Tolchinsky, 2018), dem Leseverständnis (Ahmed et al., 2014; Berninger, Abbott, Abbott, Graham & Richards, 2002; Shanahan, 2006), grammatischen Fähigkeiten (Drijbooms, Groen, & Verhoeven, 2015), dem Arbeitsgedächtnis (Drijbooms, Groen, & Verhoeven, 2017; Kellogg, 1996; Kellogg, 2008; Olive, Kellog & Piolat, 2008) und den Exekutivfunktionen (Drijbooms et al., 2015; Drijbooms et al., 2017; Olive et al., 2008) aus der Literatur bekannt. Allerdings werden angesichts der Spezifität der Aufgabe zur Informationsdichte nur maximal mittlere Korrelationen zwischen dem Testscore und den genannten Variablen erwartet. Die Korrelation zur Intelligenz kann aufgrund der im theoretischen Teil beschriebenen Anforderungen (z.B. Planungsprozesse) an Schreiber*innen nicht ausgeschlossen werden. Es werden daher mittlere Korrelationen zwischen dem Testscore und der Intelligenz erwartet.

3) Abhängigkeit des Testscores von der Klassenstufe: Zu erwarten ist ein Anstieg des Testscores in allen Schwierigkeitsstufen über die Klassenstufen (entsprechend Bereiter & Scardamalia, 1987).

7.3 Methode

7.3.1. Stichprobe

Die Stichprobe besteht aus insgesamt 541 Kindern und Jugendlichen (49,4 % männlich) der 4. bis 9. Klassenstufe (N = 41) aus Grund- und Gesamtschulen in Köln und dem Umland von Köln. In der Gesamtgruppe aller 541 Schüler*innen sprechen 75,5% Deutsch als Muttersprache. Die ausgewählten Schulen sind bezüglich sozio-ökonomischer Bedingungen möglichst homogen; es handelt sich um Schulen in innerstädtischen und ländlichen Randregionen von Köln. Eine Kontrolle der sozio-ökonomischen Bedingungen ist aus Datenschutzgründen in dieser ersten Untersuchung nicht möglich. Die Erhebung fand von April 2016 bis Juni 2017 in ruhigen Räumen an den Schulen in Gruppentestungen statt.

7.3.2 Instrumente

7.3.2.1 Die Aufgabe „Verrückte Gebäude“: eine Skala zum Informationsmanagement

Im Rahmen der Aufgabe waren die TP dazu aufgefordert, eine Bauanleitung zu einem Gebäude aus Bausteinen zu schreiben (siehe oben). Die Positionierung und Ausrichtung der Bausteine (zueinander) sollte so instruiert werden, dass ein unwissender Adressat entlang der verfassten Anleitung dazu in der Lage ist, das Gebäude fehlerfrei nachzubauen. Für die Fertigstellung der Aufgaben gab es keine Zeitbeschränkung; die Schüler*innen der Stichprobe benötigten maximal 35 Minuten für das Verfassen der Bauanleitungen zu den fünf vorgegebenen Bausteingebäuden.

Entsprechend der Aufgabenstellung ist für die Auswertung die Frage danach maßgeblich, ob die Informationen, die für den (korrekten) Nachbau notwendig sind, von der TP vollständig gegeben werden (d.h. ob die Informationsdichte adäquat ist). Allerdings ist es schwer, die notwendige Informationsdichte a priori „theoretisch“ zu bestimmen, da manche Informationen durchaus verzichtbar sind (z.B. „Stelle den Stein hochkant auf die Tischplatte“ – „hochkant“ und „Tischplatte“ möglicherweise verzichtbar), andere aber unabdingbar genannt werden müssen. Aus diesem Grund wurde anhand experimenteller Untersuchungen ein Kriterienkatalog entwickelt, der für jede Schwierigkeitsstufe die unbedingt notwendigen Informationen vorgibt.

Zur Ermittlung der unbedingt notwendigen Informationen wurden erwachsenen TP Bauanleitungen schrittweise und in Form von einzelnen kurzen Sätzen durch eine(n) Versuchsleiter*in vorgelesen. Gleichzeitig hatten die TP die Möglichkeit, die Bauanleitung selbst zu lesen. Ein so kleinschrittiges Vorgehen erschien notwendig, da kleinere Voruntersuchungen gezeigt hatten, dass TP beim eigenständigen Lesen der Bauanleitungen immer wieder wichtige Informationen überlesen hatten. Im Rahmen dieser Experimente wurde die in den Bauanleitungen gegebene Informationsdichte systematisch variiert, indem einzelne Informationen weggelassen wurden. Es wurde dann jeweils geprüft, ob der Nachbau des Bausteingebäudes trotz geringerer Informationsdichte noch möglich war. Kriterium war, dass 80 Prozent der TP das Gebäude anhand der gegebenen Informationen fehlerfrei nachbauen konnten.

Die Experimente wurden mit Studierenden der Universität zu Köln (N= 193, durchschnittliches Alter 22 Jahre, 85.5% weiblich) durchgeführt[6]. Die Anzahl der als notwendig identifizierten Informationen variierte zwischen den Komplexitätsstufen. Sie nahm wie erwartet mit steigender Komplexität zu.

Auf Basis dieses Kriterienkatalogs wurde bei der späteren Auswertung der von den Schüler*innen verfassten Texte ein Punkt für jede korrekt gegebene Information vergeben und der Quotient aus den von der/dem Schüler*in gegebenen und allen nötigen Informationen als Maß für die Informationsdichte bestimmt.

7.3.2.2 Zusätzliche Instrumente

Die sprachfreie Intelligenz wurde mit dem Grundintelligenztest (CFT 20-R; Weiß 2006) erhoben; die Raumvorstellung mit dem Mental Rotation Test (MRT; Peters, Laeng, Latham, Jackson, Zaiyouna & Richardson, 1995). Das basale Lesen wurde mit dem Salzburger Lesescreening (SLS 2-9; Wimmer & Mayringer, 2014) erfasst, der Wortschatz mit dem Wortschatz-Ergänzungstest aus dem CFT 20-R (WS-R) und über drei Subtests (Synonyme, Antonyme und Wortendungen) des Allgemeinen Deutschen Sprachtests (ADST; Steinert, 2011). Das Leseverständnis mit dem Lesegeschwindigkeits- und -verständnistest (LGVT 5-12+; Schneider, Schlagmüller & Ennemoser, 2017). Zur Messung des (phonologischen) Arbeitsgedächtnisses wurde ein Verfahren in Anlehnung an den Reading Span Test (RST; Carroll, Meis, Schulte, Vormann, Kießling & Meister, 2015) konstruiert. Hier hatten die TP die Aufgabe, eine Serie von Sätzen nacheinander zu lesen. Unmittelbar nach jedem Satz sollten die TP beurteilen, ob der jeweilige Satz eine richtige oder eine falsche Aussage trifft. Zusätzlich war es Aufgabe der TP, sich das jeweils letzte Wort des Satzes zu merken. Unmittelbar im Anschluss an eine präsentierte Satzserie (zwischen 2 bis 5 Sätze umfassend) sollten die TP dann die jeweils letzten Wörter der präsentierten Sätze entsprechend der vorgegebenen Reihenfolge schriftlich reproduzieren. Zur Messung des (räumlich-visuellen) Arbeitsgedächtnisses wurden zwei Aufgaben in Anlehnung an Hasselhorn et al. (2012) eingesetzt: Im Matrizentest wurde auf einem 4 x 4 Raster ein schwarz-weißes Muster mit ansteigender Komplexität für 2-7 Sekunden (je nach Schwierigkeit) präsentiert. Die TP mussten dieses Muster nachzeichnen. Bei der Corsiblock-

[6] Aufgrund der Tatsache, dass die spezifizierten Adressaten der Bauanleitungen Kinder bzw. Jugendliche waren, wurden die Kriterien des Bewertungsrasters 47 Schüler*innen der vierten Klasse vorgegeben. Der Versuchsablauf war der gleiche wie bei den Erwachsenen. Dieser Schritt erschien notwendig, um die Validität des Kriterienkatalogs zu überprüfen. Die gewonnenen Ergebnisse zeigen, dass die Bauergebnisse der Kinder mit denen der Erwachsenen vergleichbar sind.

Aufgabe wurden rasch nacheinander Smileys in einer unsystematischen Reihenfolge auf einer 9 x 9 Matrix präsentiert, die von den TP reproduziert werden musste. Zur Erfassung der zentralen Exekutive wurde der Star Counting Test (SCT; De Jong & Das-Smaal, 1990) durchgeführt.

7.3.3 Statistische Analyse

Zur Bestimmung der internen Konsistenz wurde Cronbachs α berechnet, zur Bestimmung der konvergenten und divergenten Validität Korrelationen mit den zusätzlich erhobenen Variablen, für die Entwicklung über die Zeit wurde eine Varianzanalyse für Messwiederholungen durchgeführt. Die faktorielle Validität wurde mittels einer konfirmatorischen Faktorenanalyse überprüft, wobei neben ML- auch Bayes-Schätzverfahren zum Einsatz kamen.

7.4 Ergebnisse

7.4.1 Interne Konsistenz

Die relative Informationsdichte der Texte zu den fünf vorgegebenen Gebäuden ist sehr gut durch eine Dimension beschreibbar. Das entsprechende Modell ist gut angepasst ($\chi^2_{(5)}$=5,35; p=0,08; RMSEA=0,04; CFI=0,99; NFI=0,98; PPP=0,49; siehe Tabelle 6). Cronbachs Alpha beträgt α=0,79. Die Aufgabenschwierigkeit steigt monoton mit den a priori festgelegten Schwierigkeitsstufen ($F_{(1, 505)}$=1250,8.; p<0,001, η_p^2=0,71).

Tabelle 6: Ergebnisse einer konfirmatorischen Faktorenanalyse zur Dimensionalität der Aufgabe „Verrückte Gebäude“

	Schätzung (S.E.)	95% CI	Schiefe	Exzess	stand. ML-Schätzung	rel. ID
Gebäude 1	1				0,53	0,82
Gebäude 2	1,43 (0,006)	[1,20; 1,75]	0,63	0,88	0,73	0,59
Gebäude 3	1,03 (0,004)	[0,84; 1,28]	0,72	1,47	0,60	0,57
Gebäude 4	1,30 (0,006)	[1,07; 1,56]	0,73	1,06	0,72	0,50
Gebäude 5	1,16 (0,005)	[0,95; 1,42]	0,55	0,61	0,68	0,43

Anmerkung: Verwendung von Bayes- und ML-Schätzer. Letzte Spalte: Mittelwert der relativen Informationsdichte (rel. ID) für die jeweilige Aufgabe

7.4.2 Konstruktvalidität / divergente und konvergente Validität

Wie erwartet, korreliert der Gesamtscore nicht mit der Raumvorstellung (MRT; r=0,07; p>0,05). Die Korrelationen mit basalem Lesen (SLS; r=0,40; p<0,01), Wortschatz (ADST Subskalen + WS-R; r=0,50; p<0,01), Grammatik (ADST Subskala; r=0,45; p<0,01), Leseverständnis (LGVT; r=0,41; p<0,01), Arbeitsgedächtnis (Faktorenscore aus Corsiblock, Matrix und RST; r=0,03; p>0,05) und zentraler Exekutive (SCT; r=0,16; p<0,01) sind von maximal mittlerer Größe. Gleiches gilt für die Intelligenz (CFT; r=0,42; p<0,01).

7.4.3 Abhängigkeit der gegebenen Informationsdichte von der Klassenstufe

Erwartet wurde ein Ansteigen der Testscores über die Klassenstufen. Eine ANOVA mit dem Messwiederholungsfaktor Klassenstufe ergibt einen signifikanten Effekt ($F_{(5,\ 536)}$=25,5; p<0,001, η_p^2=0,19). Es kommt zu einem Anstieg des Informationsgehalts der Instruktionen in den höheren Klassenstufen. Post-hoc Analysen zeigen allerdings, dass sich nur die Klassenstufen 4 von 5-7 und 8-9, sowie 5-7 von 8-9 signifikant unterscheiden (Tabelle 7).

Tabelle 7: Unterschiede in der Aufgabe „Verrückte Gebäude“ in Abhängigkeit von der Klassenstufe.

				95%-Konfidenzintervall	
Jahrgang	N	M	SE	Untergrenze	Obergrenze
4 [a]	41	0,36	0,02	0,31	0,41
5 [b]	130	0,51	0,01	0,49	0,54
6 [b]	153	0,53	0,01	0,50	0,55
7 [b]	117	0,56	0,01	0,53	0,59
8 [c]	66	0,63	0,02	0,60	0,67
9 [c]	35	0,71	0,03	0,65	0,76

Anmerkung: gleiche Indizes bezeichnen homogene Subgruppen bei α = 5%

7.5 Diskussion

Das vorliegende Kapitel beschreibt die Entwicklung eines förderdiagnostisch orientierten Verfahrens zur Erfassung der Textproduktionskompetenz für den deutschen Sprachraum. Der Ansatz verbindet zwei wichtige theoretische Modelle der Schreibforschung (Hayes & Flo-

wer, 1980; Hayes, 2012) und ergänzt sie um Aspekte der Textproduktionskompetenz (im engeren Sinne). Hierbei handelt es sich um Fähigkeiten, die für konkrete Schreibaufgaben relevant sind.

Der Ansatz löst sich von dem Konzept, Schreibkompetenz indirekt über den Umweg globaler oder kriterialer Einschätzungen von längeren Schreibleistungen zu erheben, die in der Literatur als wenig reliabel bekannt sind (vgl. Birkel & Birkel, 2002; Studie 1 dieser Dissertation). Stattdessen stellt der Ansatz – wie am Beispiel des Aufgabensets „Verrückte Gebäude“ gezeigt – den Versuch dar, entlang klarer und empirisch gesicherter Bewertungsrichtlinien Subkomponenten des Schreibprozesses möglichst eindeutig zu erfassen. Der Blick dieses in erster Linie förderdiagnostischen Ansatzes ist dabei stark auf die Intervention gerichtet. Angestrebtes Ziel ist es, die Möglichkeit zu schaffen, auf die einzelnen Komponenten der Textkompetenz einzugehen und zu bestimmen, in welchen dieser Bereiche Förderbedarf besteht. Gelingt dies, können Kinder/Jugendliche gezielt trainiert werden. Die beispielhaft präsentierte Aufgabe zeigt, dass dies prinzipiell möglich ist. Die Skala „Verrückte Gebäude“ weist eine gute Konstruktvalidität auf. Das Konstrukt ist eindimensional und theoretische Vorannahmen über divergente und konvergente Validität wurden bestätigt.

Es ist selbstverständlich, dass die Beschreibung eines in Entwicklung befindlichen Instruments zahlreiche Limitationen aufweist. Derzeit befindet sich das Projekt noch in einer Phase, in der bislang lediglich Aussagen über die Eignung einer Subskala des Testverfahrens getroffen werden können. Ein wichtiger Punkt wird dabei die Frage nach der Abgrenzung der Skalen (bzw. der einzelnen Dimensionen) voneinander sein. Die Textproduktionskompetenz ist ein, wie das Modell zeigt, komplexes Konstrukt mit mannigfaltigen Wechselwirkungen. So können z.B. Komponenten des Textmusters (Gliederungselemente, Formulierungen) gleichzeitig kohärenzstiftende Maßnahmen sein, die dazu beitragen, dass die Bedürfnisse der adressierten Personen berücksichtigt und Erwartungen der Lesenden bedient werden (siehe Kapitel 7.1). Dennoch bleibt es Ziel des vorgestellten Ansatzes, die einzelnen Testaufgaben/-skalen so zu gestalten, dass die Komponenten der Textkompetenz möglichst unabhängig erfasst/bewertet werden können; auch wenn das Verfahren vorderhand nicht den Anspruch erhebt, die einzelnen Dimensionen völlig überschneidungsfrei und vollständig messen zu können. Mitgedacht wird die Möglichkeit, im Rahmen einer Aufgabe unterschiedliche Dinge zu erheben/bewerten. Derzeit wird beispielsweise geprüft, ob sich die vorgestellte Aufgabe „Verrückte Gebäude“ ebenfalls dazu eignet, anhand der produzierten Anleitungstexte auch Elemente der Textmusterreali-

sierung zu erheben bzw. zu beurteilen. Denkbar wäre eine Bewertung der Bauanleitung daraufhin, ob es sich hierbei lediglich um eine reine – wenn auch vollständige – Aneinanderreihung von Informationen handelt (z.B. „rot liegt verkehrt herum auf blau“) oder um einen instruktiven Text, der sich gezielt an die Adressat*innen wendet und der Aufgabenstellung „Bauanleitung“ entspricht (z.B. „Lege den roten Stein verkehrt herum auf den blauen Stein“). Es bleibt jedoch zu untersuchen, inwieweit auf diese Weise (relativ) unabhängige Konstrukte gemessen werden können.

Ein weiterer wichtiger ausstehender Punkt im Kontext der Validierung ist die Bestimmung des Zusammenhangs der entwickelten Einzelskalen mit einem globalen Maß der Textproduktionskompetenz. Die Erfassung einer solchen Leistung soll über einen längeren Text erfolgen. Hierzu wurde bereits eine Schreibaufgabe entwickelt. Im Rahmen dieser Aufgabe sind TP dazu aufgefordert, die Funktionsweise eines technischen Apparates (einer Seifenblasenpistole) zu erklären. Auch hier werden die TP im Vorfeld mit dem notwendigen Vorwissen ausgestattet, indem sie unmittelbar vor der Aufforderung zur Textproduktion ein Video sehen, in welchem die Funktionsweise der Seifenblasenpistole nonverbal (mit dem Ziel, keine sprachlichen Formulierungen vorzugeben) erklärt wird. Über eine globale Bewertung des dazu entstandenen Erklärungstextes soll dann ein Globalmaß für die Textproduktionskompetenz beziehungsweise das Endprodukt eines Textproduktionsprozesses generiert werden. Wichtig ist jedoch mit Blick auf das bereits benannte Problem der fehlenden Reliabilität bei der globalen Textbewertung (siehe Kapitel 5), dass das generierte Maß für die Schreibkompetenz gewissen statistischen Reliabilitätsanforderungen genügt. Nur wenn dies der Fall ist, kann ein solches Maß zur Bestimmung von Zusammenhängen im Kontext von Validierungsuntersuchungen genutzt werden. Zur Bestimmung eines reliablen Maßes für die Textqualität soll die Methode des *Comparative Judgment* (CJ; Thurstone, 1927) genutzt werden (eine genauere Erklärung dieser Methode findet sich in Kapitel 5.4.3.1 dieser Arbeit). Der hierbei ermittelte Logit-Wert pro Text wird als Qualitätsparameter des jeweiligen Textes herangezogen. Die Nutzung dieser Werte als zuverlässige Indikatoren der Textqualität lässt sich damit legitimieren, dass die interne Konsistenz (Reliabilität) der geschätzten Logit-Werte-Verteilung über die sogenannte *Scale-Seperation-Reliabilität* (analog zu Crobachs Alpha) ermittelt werden kann (Jones & Karadeniz, 2016) und unter Einsatz der Software D-PAC (siehe ebenfalls Kapitel 5.4.3.1) vorab bestimmt werden kann, welchen Reliabilitätswert die Rangfolge der Texte mindestens erreichen soll. Die Zuverlässigkeit der globalen Textqualitätsparameter kann somit abgesichert werden. Der auf diese Weise ermittelte Wert für die globale Qualität eines von einer TP produzierten Textes kann dann mit den einzelnen Subtestergebnissen derselben Person korreliert werden. Fragen zum Zusammenhang der

entwickelten Einzelskalen und der globalen Textqualität (als Endergebnis des Schreibprozesses) können auf diese Weise zuverlässig beantwortet werden.

Dies ist jedoch nur dann der Fall, wenn auch die Ergebnisse der Einzelskalen zuverlässig sind, d.h. die im Rahmen der Testkonstruktion entwickelten Bewertungsraster der jeweiligen Subtests zu objektiven und reliablen Beurteilungen der entsprechenden Teilkomponenten der TPK im engeren Sinne führen. Die Konstruktion entsprechender Bewertungsraster ist mit besonderen Herausforderungen verbunden. Anhand der hier präsentierten Aufgabe der „Verrückten Gebäude“ lässt sich zeigen, worin Herausforderungen liegen, wenn es um die Bewertung kürzerer Texte entlang vorgegebener Kriterien (hier das Vorliegen/Nichtvorliegen notwendiger Informationen) geht. Die vorgegebenen Kriterien (beispielsweise für das erste der Bausteingebäude „grün auf blau“ und „grün mit den Füßen nach oben“) bilden auf inhaltlicher Ebene unbedingt erforderliche Informationen ab, die im Text enthalten sein müssen. Schreiber*innen haben jedoch verschiedene Möglichkeiten, diese zu verschriftlichen, d.h. es bestehen verschiedene Formulierungsmöglichkeiten, die zur Übersetzung dieser Informationen genutzt werden können. Für das Kriterium „grün mit den Füßen nach oben“ finden sich beispielsweise folgende Formulierungen: „die Brücke liegt auf dem Rücken“, „die Brücke falsch herum“, „den grünen wie eine Skaterbahn“, „der grüne mit der gebogenen Seite nach oben“ und „der grüne mit der Kante nach unten“.

Die vorliegenden Daten zeigen, dass die Varianz an möglichen Formulierungen groß ist und die Kategorisierung dieser Formulierungen als *passend* oder *unpassend* im Hinblick auf die vorgegebenen Kriterien ist nicht immer trivial. Ein Auswertungsraster, welches die objektive Bewertung der Texte mit Blick auf die vorgegebenen Informationen ermöglichen soll, muss daher eindeutige Informationen dazu enthalten, welche Formulierungsalternativen zulässig sind und welche nicht. Die Auswertung der vorliegenden Daten durch zwei unabhängige Bewerter*innen hat gezeigt, wie wichtig das Vorhandensein eindeutiger Vorgaben hierzu ist und wie groß die Unterschiede in der Bewertung ausfallen, wenn zwei Personen unabhängig voneinander – ohne entsprechende Vorgaben – für dieselben Texte entscheiden, ob bestimmte Informationen in den Texten vorhanden sind oder nicht. Besonders sprachlich kreative Übersetzungen (z.B. die Nutzung von Vergleichen oder die Nutzung figürlicher Beschreibungen) führten zu Uneinigkeiten in den Urteilen. Häufig konnte in solchen Fällen sogar durch Diskussionen der beiden Beurteiler*innen sowie ihren Austausch mit der Autorin dieser Arbeit nur schwer oder überhaupt keine Einigung zur Angemessenheit verschiedener Formulierungen gefunden werden.

Aus diesem Grund wurde eine Befragungsstudie durchgeführt, in der 98 Versuchspersonen (Lehramtsstudierende der Universität zu Köln) dazu aufgefordert wurden, die Adäquatheit bestimmter Formulierungen (N=106) mit Blick auf die zu gebende Information zu bewerten. Hierzu wurden jeder Person zwischen 43 und 49 der 106 nicht eindeutig zu bewertenden Formulierungen präsentiert. Zusätzlich hierzu sahen die Versuchspersonen ein Foto des betreffenden Bausteins in seiner Zielposition sowie das Kriterium des Auswertungsrasters (z.B. „grün mit den Füßen nach oben“). Die Versuchspersonen sollten dann entscheiden, ob die vorgegebene Formulierung ihrer Meinung nach *passend* war, d.h. zum gewünschten Bauergebnis führen würde oder nicht. Jede der Formulierungen wurde im Durchschnitt 33-mal auf diese Weise bewertet. Bestätigten mehr als 75% der zur jeweiligen Formulierung befragten Personen deren Funktionstüchtigkeit, wurde diese als *passende* Formulierung in das Bewertungsraster aufgenommen. Die übrigen Formulierungen wurden als *unpassende* Formulierungen im Bewertungsraster klassifiziert.

Ein weiterer Faktor, welchen es mit Blick auf die Objektivität der Bewertungen der Bauanleitungen zu beachten gilt, ist die in Kapitel 7.1.1.3 beschriebene Möglichkeit, dass Informationen implizit in Texten vorhanden sein können und (kompetente) Leser*innen sich diese (auf Basis ihres Vorwissens und ihrer Erwartungen an den Schreibenden) logisch erschließen können. Die vorliegenden Daten zeigen, dass sich in einigen der Bauanleitungstexte implizite Informationen finden lassen. Beispielsweise kann durch die korrekte Positionierung zweier Steine die Ausrichtung eines dritten Steins implizit vorgegeben werden. Es bedarf dann keiner expliziten Instruktion mehr zur Ausrichtung des dritten Steines. Stoßen Bewertende auf solche Instruktionen bzw. Textstellen, spielen die individuellen Vorannahmen und die hiermit verbundenen Leseverständnisprozesse der Beurteilenden eine entscheidende Rolle. Die Bewertung dieser impliziten Informationen erfolgt dementsprechend entlang subjektiver Annahmen und Verständnisprozesse. Zur Vermeidung solcher subjektiven Prozesse wurde auch der Umgang mit impliziten Informationen im Bewertungsraster klar geregelt. Für jedes der Gebäude wurde dazu in einem ersten Schritt erarbeitet, an welchen Stellen Schreiber*innen (theoretisch) implizit Informationen geben könnten. Das Bewertungsraster wurde entsprechend der entwickelten Möglichkeiten angepasst beziehungsweise ergänzt. In einigen Fällen wurden explizite Kriterien, die einander „doppelten“, aus dem Raster herausgenommen (z.B. „blau unten“ und „grün darauf“ wird zu „grün auf blau“, denn anhand dieser Information wird deutlich, dass der blaue Stein unter dem grünen sein muss. In anderen Fällen wurden „Wenn-Dann-Bedingungen“ in das Bewertungsraster aufgenommen, die angeben, unter welchen Bedingungen eine Information als implizit gegeben und damit als „im Text vorhanden“ gewertet werden muss (z.B. kann

die Information, dass zwei Steine „direkt aneinander“ stehen implizit gegeben sein, wenn die Ausrichtung eines dritten Steines, der auf den beiden Steinen liegt, so explizit und detailliert beschrieben ist, dass dieser nur dann entsprechend der Instruktion verbaut werden kann, wenn die beiden unteren Steine „direkt aneinander“ stehen).

Die ergriffenen Maßnahmen (Klassifizierung der Formulierungen als *passend/unpassend* sowie Vorgaben zum Umgang mit impliziten Kriterien) führten dazu, dass die Bewertungen der beiden Bewerter*innen ein hohes Maß an Übereinstimmung aufwiesen und bei fehlender Übereinstimmung schnell eine eindeutige Lösung entlang des Bewertungsrasters gefunden wurde. Inwiefern sich das nun vorhandene Auswertungsraster für den praktischen Einsatz eignet, d.h. inwiefern Testanwender*innen anhand des Rasters zu übereinstimmenden Beurteilungen gelangen können, gilt es zukünftig anhand einer größeren Stichprobe zu überprüfen. Kleinere Voruntersuchungen mit 48 Lehrkräften (83% weiblich) für das Unterrichtsfach Deutsch, deren durchschnittliches Alter bei 45,6 Jahren (*SD*=13,0) lag und die im Mittel über 17 Jahre (*SD*=12,1) Berufserfahrung verfügten, weisen jedoch bereits auf zwei mögliche Probleme hin: Zum einen scheint es einigen Lehrkräften schwer zu fallen, sich an die vorgegebenen Kriterien des Auswertungsrasters zu halten. Trotzdem die in den Texten vorhandenen Formulierungen eindeutig als *passend* bzw. *nichtpassend* im Bewertungsraster klassifiziert sind, weichen Lehrkräfte bei der Bewertung von diesen Vorgaben ab. So geben beispielsweise immerhin 8% der Befragten entlang der Formulierung „der grüne sieht aus wie eine Skaterrampe“ an, dass die Information, dass „der grüne mit den Füßen nach oben“ liegt, nicht im Text gegeben sei, obwohl sogar in der vorher gezeigten Beispielbewertung die gleiche Formulierung verwendet und als *passend* bewertet wurde.

Besonders schwierig scheint die Bewertung von Informationen dann, wenn zwei Informationen in einem Satz gegeben werden. Bei der Formulierung „der rote wird neben den grünen gestellt“ geben 18,8% der Befragten an, dass die Information „der rote Stein steht“ nicht gegeben sei, während anhand dieser Formulierung für 58% der Befragten die Information „der grüne Stein steht“ vorhanden ist. Ein weiteres Beispiel dafür, dass die richtige Bewertung dann besonders schwerfällt, wenn mehrere Informationen gemeinsam in einem Satz gegeben werden, findet sich bei der Bewertung folgender Instruktion: „Lege den grünen hin und stelle den blauen und den roten Stein oben drauf“. Basierend auf dieser Formulierung geben 45% der Befragten an, die Information „blau steht“ sei nicht vorhanden, allerdings meinen 10,4% der Befragten, dass hieraus hervorgehe, dass der rote Stein „rechts/links außen“ auf dem grünen stehe.

Auf Basis der dargestellten Beispiele stellt sich an dieser Stelle die Frage, welchen Einfluss syntaktische Strukturen und deren Komplexität auf die Bewertung der Bauanleitungstexte nehmen. So finden sich noch deutlich komplexere Formulierungen in den gesammelten Schüler*innentexten als die in der beschriebenen Vorstudie zur Bewertung vorgegebenen. Aber auch wenn dies eine recht spannende Forschungsfrage ist, bleibt die Anforderung bestehen, dass Personen zur Auswertung und Interpretation der Testergebnisse eines Schreibkompetenztests über entsprechende Lesekompetenzen und fachliche (sowie diagnostische) Expertise zur Einordnung der Testergebnisse verfügen müssen. Unter der Annahme, dass die unterschiedlichen Bewertungsraster und deren Anwendung im Testmanual ausführlich und eindeutig beschrieben werden, besteht im Fall der Schreibtests jedoch die Erwartung, dass Experten existieren, welche die Bewertung der Textproduktionskompetenzen im engeren Sinne entlang der vorgegeben Auswertungsraster zuverlässig vornehmen können.

8. Ausblick

Studie 1 dieser Arbeit hat gezeigt, dass weiterhin die Frage im Raum steht, ob und wie Experten der Textbewertung identifiziert werden können. Diese Frage ist – trotz des in der Entwicklung befindlichen Testverfahrens – weiterhin von Relevanz, denn es besteht die Annahme, dass das Konvergenzmodell auf Ebene der TPK im engeren Sinne noch nicht vollständig ist, beziehungsweise das Konstrukt der Textproduktionskompetenz bisher nicht in vollem Umfang erforscht und definiert wurde. Experten der Textbewertung könnten durch ihre (impliziten) Kenntnisse zu entscheidenden Faktoren und Kriterien guter Textprodukte (z.B. Kaufman et al., 2013) einen wichtigen Beitrag zur Identifikation und Operationalisierung solcher weiteren Faktoren leisten. Mit Blick auf die Ergebnisse aus Studie 1 muss jedoch festgestellt werden, dass keine der im Rahmen dieser Dissertation untersuchten (Experten-)Gruppen nach den vordefinierten Kriterien eine objektive Expertise bei der Textbewertung aufweist. Allerdings (und hierin liegt ein Nachteil der MFRA) lässt sich der prozentuale Anteil der Bewerter*innen, die eine gute Modellpassung aufweisen, d.h. konsistent bewerten, nicht nach eindeutigen Maßstäben (z.B. als hoch oder niedrig) klassifizieren.

Ergänzend zu der durchgeführten Multi-Facetten-Rasch-Analyse wurden daher Ansätze der klassischen Testtheorie genutzt, um zusätzliche Informationen zur Konsistenz (und Strenge) der untersuchten Gruppen zu generieren. Dieser zusätzliche, aber nicht alleinige Einsatz von Methoden der klassischen Testtheorie entspricht den in der Literatur gegebenen Empfehlungen

(z.B. Harting & Frey, 2013). Die Ergebnisse zeigen, dass sich bezogen auf das erste Kriterium für Expertise (Strenge bei der Beurteilung) im Rahmen einer Varianzanalyse (ANOVA) keine signifikanten Gruppenunterschiede hinsichtlich der im Mittel vergebenen Punkte bei der Bewertung der globalen Textqualität ($F_{(3,171)}$=0,693; p=0,56) ergeben, d.h. keine der untersuchten (Experten-)Gruppen war (im Vergleich zu Novizen) besonders streng bei der Textbewertung. Mit Blick auf das zweite Kriterium, die Konsistenz bei der Textbewertung, zeigen die gruppenweise ermittelten Koeffizienten der Intraklassenkorrelation, dass Novizen in ihren Bewertungen lediglich eine geringe (Cinccetti, 1994) Übereinstimmung aufweisen und die Übereinstimmungen innerhalb der beiden Expertengruppen sowie unter Lehrkräften als moderat (Cinccetti, 1994) einzuordnen sind. Demnach weist keine der Gruppen eine zufriedenstellende Beurteilerübereinstimmung auf – auch wenn sich signifikante Unterschiede hinsichtlich der Konsistenz zwischen der Gruppe der Wissenschaftler*innen und der Gruppe der Novizen ergeben. Ähnliche Befunde zeigen sich in Bezug auf die Konsistenz hinsichtlich der vorgenommenen Gewichtungen der Textbewertungskriterien (3. Kriterium von Expertise). Ein Vergleich der untersuchten Gruppen zeigt, dass sowohl die Gruppen der Wissenschaftler*innen als auch die Gruppe der Lehrkräfte ein signifikant höheres Maß an Konsistenz bei der Gewichtung der Textbewertungskriterien aufweist als die Gruppe der Novizen. Jedoch lässt sich die Konsistenz (in Form der Intraklassenkorrelation) innerhalb der Gruppe der Wissenschaftler*innen lediglich als moderat (Cincetti, 1994) klassifizieren; innerhalb der Gruppe der Lehrkräfte ist die Beurteilerübereinstimmung sogar nur gering (Cincetti, 1994).

Zusammenfassend zeigen somit auch die Befunde auf Basis der klassischen Testtheorie (für eine ausführlichere Darstellung siehe Anhang C), dass keine der untersuchten Gruppen über ein ausreichendes Maß an Expertise bei der Bewertung der Textqualität verfügt, Wissenschaftler*innen sich aber als einzige Gruppe im Hinblick auf zwei der drei Kriterien signifikant von Novizen unterscheiden. In Kombination mit den Ergebnissen der MFRA[7] legen diese Befunde nahe, die Gruppe der Wissenschaftler*innen noch einmal genauer in den Blick zu nehmen und hier nach möglichen Experten der Textbewertung zu suchen. Eine Möglichkeit hierzu stellt die Weiterverarbeitung der anhand der MFRA gewonnenen Mean-Square-Fehlerstatistiken der einzelnen Beurteiler*innen dar. Anhand dieser könnten gezielt solche Wissenschaftler*innen identifiziert werden, die konsistente Urteile bei der Textbewertung abgeben und gleichzeitig hin-

[7] Auch hier war der Anteil an Wissenschaftler*innen mit konsistenten Textbewertungen und Gewichtungen der Bewertungskriterien recht hoch. Außerdem wiesen Wissenschaftler*innen bei durchschnittlicher Gewichtung der Kriterien mehr Ähnlichkeiten auf als die übrigen Gruppen.

sichtlich der vorgenommenen Gewichtungen der Textbewertungskriterien miteinander übereinstimmen. Weisen die auf diese Weise ausgewählten Personen auch noch eine gewisse Strenge bei der Bewertung der Textqualität auf, so ließe sich gegebenenfalls doch noch eine Personengruppe mit entsprechender Expertise identifizieren. Ein ähnliches Vorgehen ist für die übrigen untersuchten Gruppen ebenfalls denkbar. Merkmale, die Personen mit entsprechender Expertise vereinen (z.B. langjährige Berufserfahrung oder Ähnliches), ließen sich auf diese Weise ebenfalls bestimmen.

Bei der Suche nach weiteren Komponenten der TPK sowie manifesten Kriterien zu deren Bewertung erscheint es außerdem sinnvoll, den Blick nicht ausschließlich auf das Endprodukt des Produktionsprozesses (den Text) und dessen Bewertung zu richten, sondern auch den Textproduktionsprozess selbst genauer in den Blick zu nehmen. Experimentelle Ansätze, welche professionelle Schreiber*innen, aber auch Schreibnovizen und „durchschnittliche“ Schreiber*innen gezielt mit Problemen und Herausforderungen des Textproduktionsprozesses konfrontieren und im Rahmen derer sich die Reaktionen und Problemlösestrategien der Schreibenden beobachten lassen, stellen einen vielversprechenden Ansatz zur Erweiterung der durch Hayes & Flower (1980) gewonnenen Erkenntnisse zum Schreibprozess dar. Kompetenzen und Strategien, die es zur erfolgreichen Bewältigung des Problemlöseprozesses beim Schreiben eines Textes braucht (und deren Zusammenspiel), ließen sich auf diese Weise beobachten und anhand konkreter Verhaltensweisen operationalisieren. Die Idee zur Definition von Mindeststandards/Minimalanforderungen für funktionale Texte bzw. deren Produktion könnte anhand eines Vergleiches von Gruppen mit unterschiedlicher Schreibexpertise weiterverfolgt werden.

Eine letzte gestellte, aber zum jetzigen Zeitpunkt noch nicht beantwortete Frage ist die danach, welchen Einfluss die verschiedenen Komponenten der TPK im engeren Sinne auf die globale Textqualität beziehungsweise das Gesamtkonstrukt der Textproduktionskompetenz nehmen. Methodisch gesprochen stellt sich die Frage nach dem Anteil der Varianz, welchen die postulierten Dimensionen der TPK im engeren Sinne an der Textqualität bzw. dem Gesamtkonstrukt der Textproduktionskompetenz erklären können. In den Diskussionskapiteln beider hier vorgestellten Studien wird diese Frage aufgeworfen, jedoch unterscheidet sich die jeweilige Perspektive hierauf. Während Studie 1 die Bewertenden (verschiedener Gruppen) und deren Urteile in den Blick nimmt und damit nach dem Anteil der Varianz am Gesamturteil zur Textqualität fragt, welcher durch die einzelnen Teildimensionen der Textqualität (Kriterien) erklärt/nicht erklärt werden kann und welches Gewicht diesen Teildimensionen hierbei jeweils

zukommt, liegt der Schwerpunkt in Studie 2 auf den Schreiber*innen und deren Textproduktionskompetenzen. Hier geht es um die Fragen danach, wie viel Varianz des Gesamturteils zur Textqualität (Textproduktionskompetenz als Gesamtkonstrukt) durch tatsächliche Schreibleistungen in Form einzelner Subtests erklärt/nicht erklärt werden kann und welches Gewicht den Teildimensionen der TPK im engeren Sinne dabei jeweils zukommt. Eine Kombination dieser beiden Perspektiven beziehungsweise der daraus gewonnenen Befunde eröffnen die Möglichkeit eines Vergleiches der Ergebnisse indirekter und direkter Messungen der TPK und damit die Untersuchungen von Fragestellungen zur Validität beider Zugänge. Daten zur Beantwortung der beiden Fragen liegen vor.

Zum Ende dieser Arbeit wird damit klar: Ähnlich wie der Schreibprozess ist auch der Forschungsprozess ein konvergenter Prozess. Die im Rahmen dieser Arbeit gewonnenen Befunde dienen als *Sprungbrett* für weitere Forschungsideen und Forschungsvorhaben.

9. Weitere Forschungsvorhaben

Neben den im vorherigen Kapitel aufgeworfenen Fragen bestehen folgende weitere Forschungsvorhaben:

- Empirische Absicherung der einzelnen Komponenten der TPK im engeren Sinne. Hierzu erfolgt eine Operationalisierung der einzelnen Komponenten in Form von Testaufgaben entlang der in Kapitel 7.1 aufgeführten Konstruktdefinitionen. Im Hinblick auf die Operationalisierung der Textproduktionskompetenz in Form von Teilkomponenten (Subtests) steht die Frage im Raum, inwiefern es möglich ist, einzelne Komponenten der TPK im engeren Sinne voneinander abzugrenzen und getrennt zu erfassen.
- Darüber hinaus sollen die Konstrukte detaillierter untersucht und genauer beschrieben werden. Auf einzelne Teildimensionen bezogen stehen hierbei folgende Forschungsfragen im Raum:
    a. Konstrukt der „Perspektivübernahme“ beziehungsweise der „Adressatenorientierung“: Wie/auf welcher Basis bauen Schreibende ein Modell zum Adressatenkreis auf? Wie wird dieses Modell im Schreibprozess konkret berücksichtigt und welche Schreibhandlungen werden mit Blick auf den Adressatenkreis vorgenommen? Sind die sprachliche und die inhaltliche Ebene im Hinblick auf den Adressaten voneinander zu trennen? Sind die Maßnahmen zur Adressatenorien-

tierung textsortenspezifisch oder über Textsorten hinweg miteinander vergleichbar? Welche Rolle spielt das eigene Leseverständnis bei der Auswahl von Maßnahmen zur Adressatenorientierung? Welcher Zusammenhang besteht zum Konstrukt der Kohärenz und wie groß sind die Überschneidungen der beiden Konstrukte?

b. Konstrukt „Textmusterwissen“ und Anwendung dieser Muster (samt zugehöriger Textprozeduren): Ab welchem Alter und wie entwickelt sich das Textmusterwissen? Welcher Zusammenhang besteht zwischen dem deklarativen Wissen um Textmuster und der Kompetenz, diese tatsächlich anwenden zu können? Welche Rolle kommt dem Textmuster bezogen auf die anderen Komponenten der Textproduktionskompetenz im engeren Sinne zu?

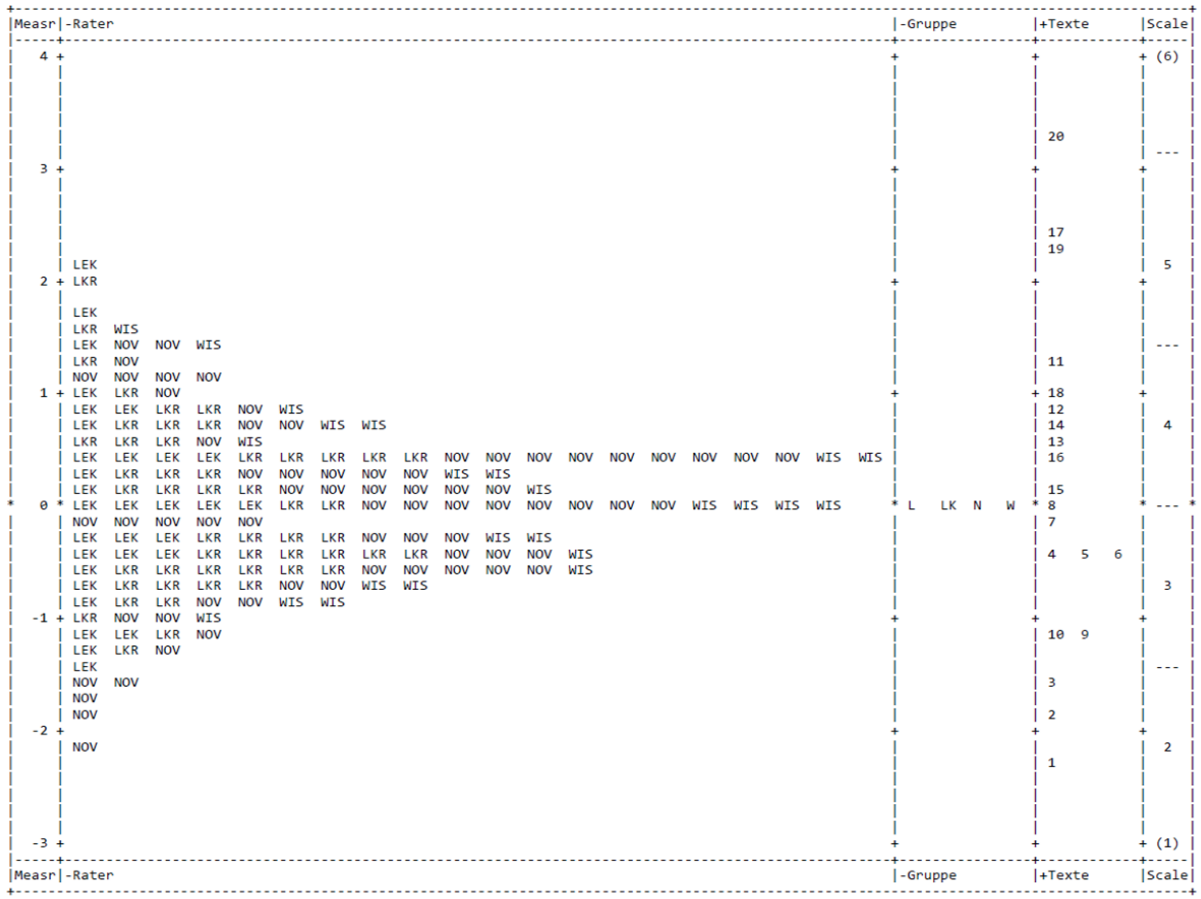
c. Konstrukt der „Kohärenz“: Welchen Anteil haben Grammatik und/oder Semantik am Zustandekommen der globalen Kohärenz? Ist die Kohärenz, ähnlich wie die Adressatenorientierung, eher ein Ergebnis verschiedener Teilkomponenten der Textproduktionskompetenz?

- Ergänzend zu bisherigen Untersuchungen zum Schreibprozess sollen die einzelnen Phasen und Prozesselemente des Schreibprozesses noch genauer untersucht werden. Konkret geht es darum, folgende Fragen zu beantworten: Welche Unterschiede ergeben sich zwischen Schreibnovizen und Schreibexperten im Hinblick auf die einzelnen Prozesselemente? Wirken sich Planungsprozesse, die vor Beginn des eigentlichen Produktionsprozesses liegen (z.B. in Form von vorab angefertigten schriftlichen Gliederungen), positiv auf den Schreibprozess bzw. die Qualität des Schreibproduktes aus? Wie stark wird der anfängliche Plan im Laufe des Schreibprozesses verändert? Wie gehen Schreiber*innen mit Planänderungen um/wie passen sie sich und den Text an diese an? Unterscheiden sich Planungsprozesse mit Blick auf die Textsorte? Wie werden aus (vagen) Ideen und Plänen gelungene Übersetzungen? Wodurch werden Überarbeitungsprozesse gesteuert? Welche Faktoren entscheiden darüber, wie tiefgreifend Überarbeitungsprozesse sind?
- Anhand der gewonnenen Daten aus den Erhebungen zur Konstruktion des psychometrischen Testverfahrens und den Untersuchungen zum Schreibprozess (Vergleich von Schreibexperten und Schreibnovizen) sollen Rückschlüsse auf die Entwicklung der Textproduktionskompetenz gezogen werden.

Anhang

Anhang A

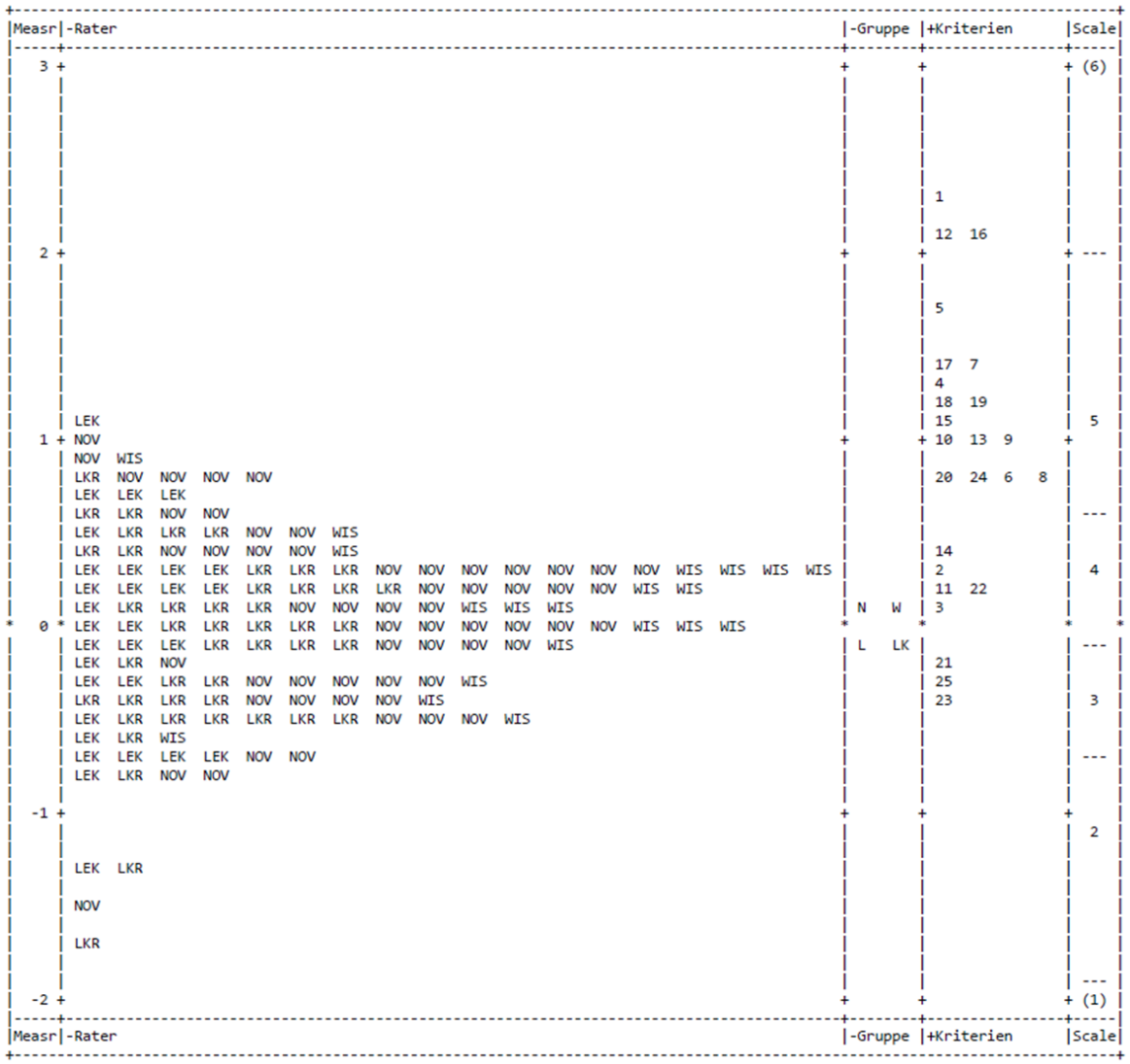
Abbildung A.1: Facettenraum der MFRA zu Model A



Anmerkung: Die erste Spalte zeigt die Logit-Skala. Die Werte auf dieser Skala zeigen die Strengemaße der Bewertenden und die globale Qualität der bewerteten Texte an. Die zweite Spalte zeigt die geschätzte Verteilung der Strengemaße der Bewertenden. Jede Abkürzung steht für eine Person (LKR = Lehrkräfte; WIS = Wissenschaftler*innen; LEK = Lektor*innen und Autor*innen; NOV = Novizen). In der dritten Spalte wurden die Gruppierungsvariablen hinzugefügt, aber nicht kalibriert, da diese als Dummy-Facette in der Analyse verwendet wurden. Die vierte Spalte zeigt die geschätzten Textqualitätsparameter. Jede Zahl steht für einen Text (1 bis 20). Die fünfte Spalte zeigt die sechsstufige Bewertungsskala auf der Logit-Skala, sodass die Textqualität entsprechend der sechsstufigen Bewertungsskala definiert werden kann. Die Schwellenwerte für die Ratingskala werden durch horizontale gestrichelte Linien dargestellt.

Anhang B

Abbildung B.1: Facettenraum der MFRA zu Model B



Anmerkung: Die erste Spalte zeigt die Logit-Skala. Die Werte auf dieser Skala zeigen die Strengemaße der Bewertenden und die globale Qualität der bewerteten Texte an. Die zweite Spalte zeigt die geschätzte Verteilung der Strengemaße der Bewertenden. Jede Abkürzung steht für eine Person (LKR = Lehrkräfte; WIS = Wissenschaftler*innen; LEK = Lektor*innen und Autor*innen; NOV = Novizen). In der dritten Spalte wurden die Gruppierungsvariablen hinzugefügt, aber nicht kalibriert, da diese als Dummy-Facette in der Analyse verwendet wurden. Die vierte Spalte zeigt die geschätzte Relevanz/das Gewicht der Bewertungskriterien. Jede Zahl steht für ein Kriterium (1 bis 25; siehe Kapitel 5.3.3). Die fünfte Spalte zeigt die sechsstufige Bewertungsskala auf der Logit-Skala, sodass die Relevanz/das Gewicht der Kriterien entsprechend der sechsstufigen Bewertungsskala definiert werden kann. Die Schwellenwerte für die Ratingskala werden durch horizontale gestrichelte Linien dargestellt.

Anhang C

Bezogen auf die Forschungsfragen in Studie 1 dieser Dissertation (Kapitel 5) wurden ergänzend zur MFRA ebenfalls Analysemethoden verwendet, die auf der klassischen Testtheorie basieren. Die Ergebnisse dieser Verfahren werden im Folgenden entlang der in Kapitel 5.1 definierten Kriterien von Expertise präsentiert.

Kriterium 1: Strenge bei der Textbewertung

Zur Überprüfung der Frage danach, ob und inwiefern sich die Strenge der untersuchten (Experten-)Gruppen bei der Textbewertung von denen der Novizen unterscheidet, wurden die im Mittel vergebenen Punkte für die Textqualität (siehe Tabelle C.1) zwischen den Gruppen verglichen. Hierzu wurde eine Varianzanalyse (ANOVA) eingesetzt. Die Ergebnisse dieser Analyse zeigen, dass sich keine signifikanten Gruppenunterschiede hinsichtlich der im Mittel vergebenen Punkte bei der Bewertung der globalen Textqualität der 20 Schüler*innentexte ergeben ($F_{(3,171)}$=0,693; p=0,56). Weder Lehrkräfte noch eine der Expertengruppen unterscheiden sich hinsichtlich ihrer Strenge von Novizen; alle Gruppen sind im Hinblick auf ihre Strenge miteinander vergleichbar, keine der Gruppen erfüllt demnach das erste Expertisekriterium.

Tabelle C.1: Mittelwert der vergebenen Punkte bei der Bewertung der globalen Textqualität im Gruppenvergleich

Lehrkräfte	Wissenschaftler*innen	Lektor*innen & Autor*innen	Novizen
72,57 (9,56)	70,17 (9,22)	70,90 (11,38)	73,13 (10,50)

Kriterium 2: Konsistenz bei der Textbewertung

Zur Berechnung der Beurteilerübereinstimmung (Konsistenz der Beurteilenden) wurde die Intraklassenkorrelation (ICC; Wirtz & Caspar, 2002) verwendet. Die Intraklassenkorrelation und das zugehörige Konfidenzintervall (95%) wurden basierend auf einem zweifaktoriellen Modell, in dem die Bewertenden als zufälliger Faktor behandelt wurden (*two-way, random*),

berechnet. Datengrundlage zur Berechnung der (absoluten) Übereinstimmung (*agreement*) zwischen den Bewertenden waren die einzelnen Textqualitätsratings einer bewertenden Person (*single measure*).

Die ICC-Maße der einzelnen Gruppen wurden anhand der folgenden Formel ermittelt:

$$ICC(2,1) = \frac{MS_{ZW} - MS_{res}}{MS_{ZW} + (k-1) \cdot MS_{res} + \frac{k}{n} \cdot (MS_{rat} - MS_{res})}$$

In der Formel steht „k“ für die Anzahl der Bewertenden (Rater), „n“ steht für die Anzahl der zu bewertenden Objekte (hier Texte). „MS_{zw}“ gibt mit (df = n-1) die Varianz zwischen den Objekten an, d.h. die Varianz der Textqualität. „MS_{res}“ mit [df = (n-1) · (k-1)] steht für die Rest- bzw. Fehlervarianz. Mit „MS_{rat}“ (mit df = k-1) wird die Varianz zwischen den Ratern, d.h. in deren Beurteilungen der Textqualität, beschrieben. Das Maß der Intraklassenkorrelation gibt folglich an, wie hoch der Anteil der Varianz in den Beurteilungen der Rater ist, der durch die Varianz zwischen den Objekten (und damit die tatsächlich vorhandene Varianz) erklärt werden kann. Je größer der Anteil der „wahren Varianz“ an der Varianz zwischen den Beurteilenden ist, desto höher fällt das Maß der ICC und damit die Reliabilität zwischen den Ratern aus.

Entlang der Formel ist das Maß der ICC abhängig von der Stichprobengröße. Da sich die untersuchten Gruppen hinsichtlich der Stichprobengröße voneinander unterscheiden, müssen für einen Vergleich der berechneten ICC Koeffizienten sowohl die ICC Koeffizienten selbst als auch die zugehörigen Konfidenzintervalle anhand der Spearman-Brown-Formel korrigiert werden (siehe de Vet, Mokking, Mosmuller & Terwee, 2017). Erst nach entsprechender Korrektur sind die ICC-Koeffizienten der einzelnen Gruppen direkt miteinander vergleichbar.

Folgende Formel wird zur Korrektur verwendet:

$$r_2 = \frac{k \cdot r_1}{1 + (k-1) \cdot r_1}$$

Hierin steht k für den Verkürzungsfaktor (hier ermittelt anhand des Verhältnisses aus der Beobachtungsanzahl der jeweils betrachteten Gruppe zu der Beobachtungsanzahl der kleinsten

Stichprobe) und r_1 für das empirisch bestimmte Reliabilitätsmaß (hier der errechnete ICC-Koeffizient). Anhand dieser beiden Werte lässt sich r_2 als korrigiertes Reliabilitätsmaß ermitteln.

Die auf die beschriebene Weise ermittelten und korrigierten Maße der Beurteilerübereinstimmung sind in Tabelle C.2 dargestellt. Eine Einordnung dieser Werte nach Cincetti (1994) zeigt, dass Novizen in ihren Bewertungen lediglich eine *geringe* (*poor*) Übereinstimmung aufweisen. Die Übereinstimmungen innerhalb der beiden Expertengruppen sowie unter Lehrkräfte lassen sich als *moderat* (*fair*) bezeichnen. Keine der Gruppen weist eine zufriedenstellende Beurteilerübereinstimmung auf.

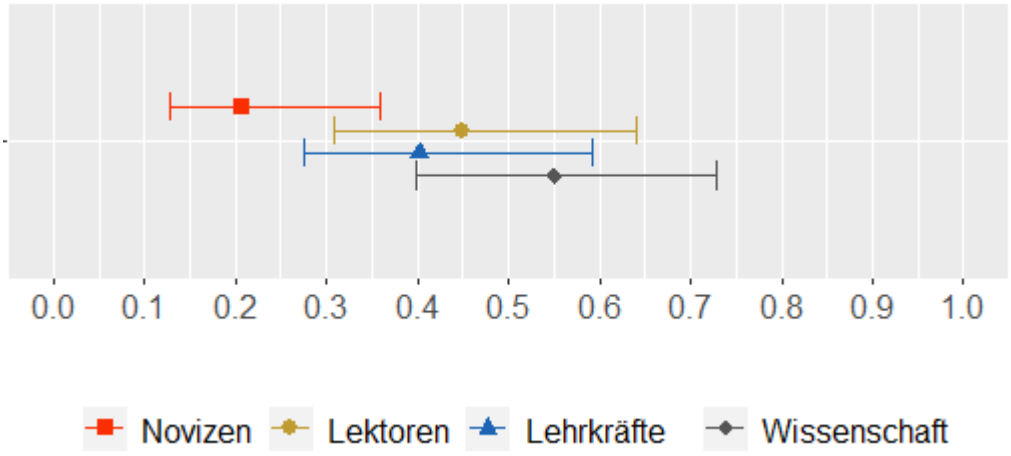
Tabelle C.2: Beurteilerübereinstimmung bei der Bewertung der globalen Textqualität im Gruppenvergleich

			95% Konfidenzintervall	
	N	ICC (2,1)	untere Grenze	obere Grenze
Lehrkräfte	51	0,40	0,21	0,60
Wissenschaftler*innen	24	0,55	0,40	0,73
Lektor*innen & Autor*innen	31	0,45	0,31	0,64
Novizen	69	0,21	0,13	0,36

Anmerkung: Die angegeben ICC-Koeffizienten sowie die Grenzen der Konfidenzintervalle wurden anhand der Spearman-Brown-Formel korrigiert.

Ein Gruppenvergleich zeigt, dass Lehrkräfte sich hinsichtlich ihrer Konsistenz bei der Bewertung der Textqualität weder von Novizen noch von den beiden Expertengruppen unterscheiden. Signifikante Unterschiede ergeben sich lediglich zwischen der Gruppe der Wissenschaftler*innen und der Gruppe der Novizen; die gebildeten Konfidenzintervalle der jeweiligen ICC-Koeffizienten überschneiden einander nicht (siehe Abbildung C.1). Die Gruppe der Wissenschaftler*innen weist demnach ein signifikant höheres Maß an Konsistenz innerhalb der vorgenommenen Textbewertungen auf als die der Novizen. Die Gruppe der Wissenschaftler*innen unterscheidet sich hinsichtlich ihrer Konsistenz jedoch nicht von der Gruppe der Lehrkräfte oder der Gruppe der Lektor*innen und Autor*innen.

Abbildung C.1: Beurteilerübereinstimmung bei der Bewertung der globalen Textqualität im Gruppenvergleich



Kriterium 3: Konsistenz bei der Gewichtung von Textbewertungskriterien

Auch zur Beantwortung dieser Forschungsfrage wurden für die einzelnen Gruppen Intraklassenkorrelationen berechnet. Das Vorgehen hierzu entspricht dem oben (unter Kriterium 2) dargestellten. Anstelle der Textbewertungen der einzelnen Rater bilden hier die vorgenommenen Gewichtungen der Textbewertungskriterien durch die einzelnen Rater die Datengrundlage. Aufgrund der Tatsache, dass die Stichprobengröße zwischen den Gruppen erneut variiert, wurde auch hier eine Korrektur der geschätzten Koeffizienten entlang der Spearman-Brown-Formel vorgenommen.

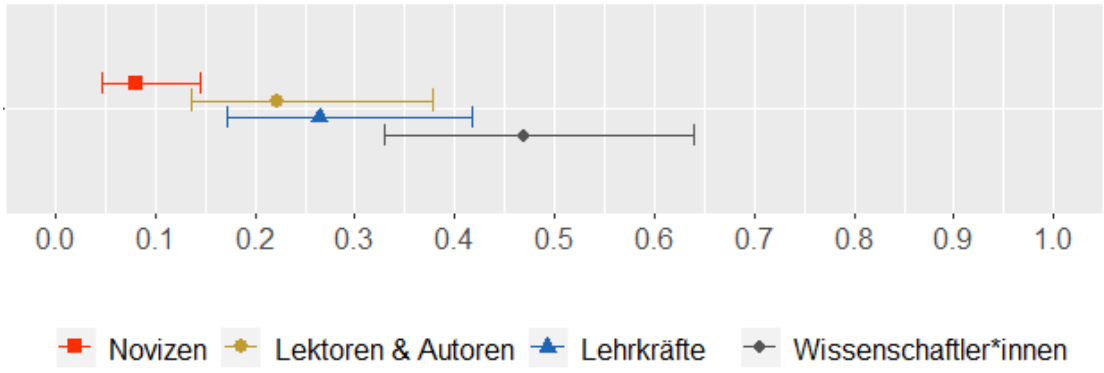
Erneut zeigt sich (siehe Tabelle C.3), dass innerhalb der Gruppe der Wissenschaftler*innen das Maß der Beurteilerübereinstimmung am höchsten ist. Der ICC-Wert lässt sich als *moderat* (*fair*; Cincetti, 1994) klassifizieren. In allen anderen Gruppen ist die Beurteilerübereinstimmung *gering* (*poor*; Cincetti, 1994); bei Novizen liegt diese sogar nur knapp über 0.

Tabelle C.3: Beurteilerübereinstimmungen bei der Gewichtung der Textbewertungskriterien im Gruppenvergleich

			95% Konfidenzintervall	
	N	ICC (2,1)	untere Grenze	obere Grenze
Lehrkräfte	45	0,26	0,17	0,42
Wissenschaftler*innen	19	0,47	0,33	0,64
Lektor*innen & Autor*innen	30	0,22	0,14	0,38
Novizen	58	0,08	0,05	0,14

Ein Vergleich der Gruppen (siehe Abbildung C.2) zeigt: Die Gruppen der Wissenschaftler*innen, Lehrkräfte und Lektor*innen und Autor*innen unterscheiden sich nicht signifikant voneinander hinsichtlich ihrer Übereinstimmungen bezüglich der vorgenommenen Gewichtungen. Es zeigen sich aber signifikante Unterschiede zwischen der Gruppe der Wissenschaftler*innen und der Gruppe der Novizen sowie zwischen der Gruppe der Lehrkräfte und der Gruppe der Novizen. Beide Gruppen, Wissenschaftler*innen und Lehrkräfte, weisen ein signifikant höheres Maß an Übereinstimmung in ihren Gewichtungen der Textbewertungskriterien auf als Novizen und gewichten diese somit konsistenter. Die Gruppe der Lektor*innen und Autor*innen unterscheidet sich nicht signifikant von der der Novizen.

Abbildung C.2: Beurteilerübereinstimmung bei der Gewichtung der Textbewertungskriterien im Gruppenvergleich



Vergleich der Befunde mit denen der MFRA

Bei einem Vergleich der hier dargestellten Befunde mit denen der MFRA in Kapitel 5.5 lassen sich Gemeinsamkeiten, aber auch Unterschiede feststellen.

Mit Blick auf Kriterium 1 (Strenge der Beurteilungen) ergeben sich zwischen den Ergebnissen der beiden angewandten Verfahren keine Unterschiede. Beide zeigen, dass es keine Gruppenunterschiede hinsichtlich der Strenge, d.h. der im Mittel vergebenen Punkte für die Textqualität gibt. Im direkten Vergleich bietet die MFRA jedoch die Möglichkeit, gleichzeitig auch etwas zu den Unterschieden in den Bewertungsmaßstäben (Strenge/Milde) innerhalb der verschiedenen Gruppen zu erfahren und diese Unterschiede (und hierin besteht der eigentliche Vorteil der Methodik) sinnvoll zu quantifizieren und zwischen den Gruppen vergleichen zu können.

Mit Blick auf Kriterium 2 lässt sich erneut feststellen, dass die Ergebnisse der beiden Verfahren (ICC vs. MFRA) zu ähnlichen Ergebnissen führen. Beide zeigten, dass die Gruppe

der Wissenschaftler*innen das höchste Maß an Konsistenz bei der Textbewertung aufweist, gefolgt von der Gruppe der Lehrkräfte. Für die Gruppe der Lektor*innen und Autor*innen sowie die Gruppe der Novizen fallen die Konsistenzmaße (deskriptiv betrachtet!) geringer aus. Entlang der Befunde der MFRA lassen sich zwischen den untersuchten Gruppen keine signifikanten Unterschiede hinsichtlich der Konsistenz bei der Textbewertung finden. Im Gegensatz dazu finden sich bei einem Vergleich der ICC-Koeffizienten signifikante Unterschiede zwischen der Gruppe der Wissenschaftler*innen und der der Novizen. Wissenschaftler*innen weisen eine signifikant höhere Übereinstimmung (Konsistenz) bei der Textbewertung auf als Novizen. Dennoch zeigen die Befunde der Intraklassenenkorrelation auch, a) dass Wissenschaftler*innen nicht konsistenter in ihren Urteilen sind als Lehrkräfte oder Lektor*innen und Autor*innen (welche sich widerum nicht von Novizen unterscheiden) und b) dass das Ausmaß an Konsistenz (der ICC) auch bei den Wissenschaftler*innen nur als „moderat“ eingeordnet und damit nicht als zufriedenstellend im Hinblick auf die Zuverlässigkeit der Textbewertungen betrachtet werden kann.

Im Hinblick auf Kriterium 3 zeigen sich die deutlichsten Unterschiede zwischen den Befunden der MFRA und den Resultaten der ICC-Berechnungen. Entlang der ICC findet sich in der Gruppe der Novizen die geringste Übereinstimmung zwischen den Bewertenden, was deren Gewichtungen der Textbewertungskriterien angeht; im Rahmen der MFRA war dies für die Gruppe der Lektor*innen und Autor*innen der Fall – hier war der Anteil der Personen, die konsistente Gewichtungen vornahmen, innerhalb der Gruppe der Lektor*innen und Autor*innen teilweise sogar signifikant geringer als innerhalb der übrigen Gruppen – vor allem dann, wenn es um die Gewichtung von Kriterien mit durchschnittlicher Relevanz ging. Im Rahmen der MFRA ließen sich jedoch – im Gegensatz zu den Befunden der ICC – keine signifikanten Unterschiede zwischen a) der Gruppe der Wissenschaftler*innen und der Gruppe der Novizen sowie b) der Gruppe der Lehrkräfte und der Gruppe der Novizen finden. Jedoch bleibt auch hier festzuhalten, dass die Konsistenz innerhalb der Gruppe der Wissenschaftler*innen lediglich moderat ist und die Gruppe der Lehrkräfte sogar nur geringe Beurteilerübereinstimmungen aufweist.

Zusammenfassend finden sich somit mehr Ähnlichkeiten als Unterschiede in den Befunden der angewandten Verfahren. Eine Triangulation der Ergebnisse weist jedoch darauf hin, dass sich die Gruppe der Wissenschaftler*innen zumindest mit Blick auf zwei der vorab definierten Kriterien doch erheblich (und entlang der klassischen Testtheorie sogar signifikant) von Novizen unterscheidet. Diese Erkenntnis bietet Anlass für weitere Forschungsbemühungen (s. Kapitel 8).

Literatur

Ahmed, Y., Wagner, R. K. & Lopez, D. (2014). Developmental relations between reading and writing at the word, sentence, and text levels: A latent change score analysis. *Journal of Educational Psychology*, 106, 419-434.

Alamargot, D., Caporossi, G., Chesnet, D. & Ros, C. (2011). What makes a skilled writer? Working memory and audience awareness during text composition. *Learning and Individual Differences*, 21(5), 505-516.

Allen, L. K., Snow, E. L. & McNamara, D. S. (2016). The narrative waltz: The role of flexibility in writing proficiency. *Journal of Educational Psychology*, 108(7), 911-924.

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5), 997-1013.

Amabile, T. M. (1996). *Creativity in Context: Update to the Social Psychology of Creativity*. Boulder, CO: Westview.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561-573.

Augst, G., Disselhoff, K., Henrich, A., Pohl, T. & Völzing, P. L. (2007). *Text-Sorten-Kompetenz. Eine echte Longitudinalstudie zur Entwicklung der Textkompetenz im Grundschulalter.* Frankfurt am Main: Peter Lang.

Averintseva-Klisch, M. (2013). *Textkohärenz*. Heidelberg: Universitätsverlag Winter.

Bach, R., Schmidt, B.M., Schabmann, A. & van Koll, S. (2016). Braucht mein Friseur wirklich Zirkel und Lineal? – Schulisches Basiswissen im Kontext der Ausbildungsreife. *Heilpädagogische Forschung*, 42, 61-72.

Bachmann, T. (2002). *Kohäsion und Kohärenz: Indikatoren für Schreibentwicklung. Zum Aufbau kohärenzstiftender Strukturen in instruktiven Texten von Kindern und Jugendlichen.* Innsbruck: Studien-Verlag.

Bachmann, T. & Becker-Mrotzek, M. (2017). Schreibkompetenz und Textproduktion modellieren. In: M. Becker-Mrotzek, J. Grabowski & T. Steinhoff (Hrsg.), *Forschungshandbuch empirische Schreibdidaktik*, 25-54. Münster, New York: Waxmann.

Baer, M., Fuchs, M., Reber-Wyss, M., Jurt, U. & Nussbaum, T. (1995). Das „Orchester-Modell“ der Textproduktion. In: J. Baurmann & R. Weingarten (Hrsg.), *Schreiben: Prozesse, Prozeduren und Produkte*, 173-201. Opladen: Westdeutscher Verlag.

Beck, S. W., Llosa, L., Black, K. & Anderson, A. T. (2018). From assessing to teaching writing: What teachers prioritize. *Assessing Writing*, 37, 68-77.

Becker-Mrotzek, M. (2003). Wie schreibt man eine Bedienungsanleitung. *Praxis Deutsch*, 179, 33-36.

Becker-Mrotzek, M. (2014). Schreibkompetenz. In: J. Grabowski (Hrsg.), *Sinn und Unsinn von Kompetenzen: Fähigkeitskonzepte im Bereich von Sprache, Medien und Kultur*, 51-72. Opladen: Verlag Barbara Budrich.

Becker-Mrotzek, M. & Böttcher, I. (2012). *Schreibkompetenz entwickeln und beurteilen: Buch mit Kopiervorlagen* (4., überarbeitete Auflage). Berlin: Cornelsen Scriptor.

Becker-Mrotzek, M., Grabowski, J., Jost, J., Knopp, M. & Linnemann, M. (2014). Adressatenorientierung und Kohärenzherstellung im Text. Zum Zusammenhang kognitiver und sprachlich realisierter Teilkomponenten von Schreibkompetenz. *Didaktik Deutsch*, 19 (37), 21-43.

Becker-Mrotzek, M. & Schindler, K. (2007). Schreibkompetenz modellieren. In: M. Becker-Mrotzek & K. Schindler (Hrsg.), *Texte schreiben*, 7-26. Kobes. Köln: Gilles & Francke.

Becker-Mrotzek, M. & Schindler, K. (2008). Schreibkompetenz modellieren, entwickeln und testen. In: Martin Böhnisch (Hrsg.), *Beiträge zum 16. Symposion Deutschdidaktik „Kompetenzen im Deutschunterricht“.* Sonderheft *Didaktik Deutsch*, 94–106.

Ben-Simon, A. & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning and Assessment*, 6(1), 1-47.

Bereiter, C. (1980). Development in writing. *Cognitive Processes in Writing*, 73-93.

Bereiter, C. & Scardamalia, M. (1987). *The Psychology of Written Composition*. New York: Routledge.

Berninger, V. W., Abbott, R. D., Abbott, S. P., Graham, S. & Richards, T. (2002). Writing and reading: Connections between language by hand and language by eye. *Journal of Learning Disabilities*, 35(1), 39-56.

Birkel, P. & Birkel, C. (2002). Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. *Psychologie in Erziehung und Unterricht*, 49, 219-224.

Birkel, P. (2003). Aufsatzbeurteilung – ein altes Problem neu untersucht. *Didaktik Deutsch*, 9(15), 46-63.

Boone, S., Thys, S., Van Avermaet, P. & Van Houtte, M. (2018). Class composition as a frame of reference for teachers? The influence of class context on teacher recommendations. *British Educational Research Journal*, 44(2), 274-293.

Böttcher, I. & Becker-Mrotzek, M. (2003). *Texte bearbeiten, bewerten und benoten. Schreibdidaktische Grundlagen und unterrichtspraktische Anregungen.* Berlin: Cornelsen.

Bradley, R. A. & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39, 324-345.

Bramley, T. (2015). *Investigating the Reliability of Adaptive Comparative Judgment*. Cambridge: Cambridge Assessment.

Bromme, R. (2008). Lehrerexpertise: Eine psychologische Konzeption für die Entwicklung und Erforschung des Wissens und Könnens von Lehrern. In: W. Schneider & M. Hasselhorn (Hrsg.), *Handbuch der pädagogischen Psychologie*, 159-167. Göttingen: Hogrefe.

Carroll, R., Meis, M., Schulte, M., Vormann, M., Kießling, J. & Meister H. (2015). Development of a German reading span test with dual task design for application in cognitive hearing research. *International Journal of Audiology*, 54, 136-141.

Castillo, C. & Tolchinsky, L. (2018). The contribution of vocabulary knowledge and semantic orthographic fluency to text quality through elementary school in Catalan. *Reading and Writing*, 31, 293-323.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284-290.

Clark, C., Hodges, D., Stephan, J. & Moldovan, D. (2005). Moving QA towards reading comprehension using context and default reasoning. *AAAI 2005 Workshop on Inference for Textual Question Answering*, 6-12. California: AAAI Press

Cooksey, R. W., Freebody, P. & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation*, 13(5), 401-434.

Crossley, S. A. & McNamara, D. S. (2016). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research*, 7, 351-370.

Cumming, A., Kantor, R. & Powers, D. E. (2001). Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework. *TOEFL Monograph Series, MS-22.* Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R. & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, 86, 67-96.

Dam-Jensen, H. & Heine, C. (2013). Writing and translation process research: Bridging the gap. *Journal of Writing Research*, 5, 89-101

De Jong, P. F. & Das-Smaal, E. A. (1990). The Star Counting Test: An attention test for children. *Personality and Individual Differences*, 11, 597-604.

de Vet, H. C., Mokkink, L. B., Mosmuller, D. G. & Terwee, C. B. (2017). Spearman-Brown prophecy formula and Cronbach's alpha: different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, 85, 45-49.

Drijbooms, E., Groen, M. A. & Verhoeven, L. (2015). The contribution of executive functions to narrative writing in fourth grade children. *Reading and Writing*, 28, 989-1011.

Drijbooms, E., Groen, M. A. & Verhoeven, L. (2017). How executive functions predict development in syntactic complexity of narrative writing in the upper elementary grades. *Reading and Writing*, 30, 209-231.

Eckes, T. (2004). Beurteilerübereinstimmung und Beurteilerstrenge. *Diagnostica*, 50(2), 65-77.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, 2(3), 197–221.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155-185.

Eckes, T. (2009). Many-facet Rasch measurement. In: V. Aryadoust & M. Raquel (Hrsg.), *Quantitative Data Analysis for Language Assessment* (Bd. I), 53-176. London, New York: Routledge.

Ehlich, K. (1983). Text und sprachliches Handeln. Die Entstehung von Texten aus dem Bedürfnis nach Überlieferung. In: S. Kammer & R. Lüdeke (Hrsg.), *Texte zur Theorie des Textes*, 228-245. Stuttgart: Reclam.

Farnan, N. & Fearn, L. (2008). Writing in the disciples: More than just writing across the curriculum. In: D. Lapp, J. Foold & N. Farnan (Hrsg.), *Content Area Reading and Learning*, 403-424. New York, NY: Erlbaum.

Feenstra, H. (2014). *Assessing Writing Ability in Primary Education: On the Evaluation of Text Quality and Text Complexity.* Enschede: Universität Twente.

Feilke, H. (2010). „Aller guten Dinge sind drei" – Überlegungen zu Textroutinen & literalen Prozeduren. In: I. Bon, T. Gloning & D. Kaltwasser (Hrsg.), Fest-Platte für Gerd Fritz. Gießen 17.05.2010. (http://www.festschrift-gerd-fritz.de/files/feilke_2010_literale-prozeduren-und-textroutinen.pdf; zuletzt abgerufen am 17.02.2020).

Flower, L. (1994). *The Construction of Negotiated Meaning: A Social Cognitive Theory of Writing*. Southern Illinois: University Press.

Flower, L. & Hayes, J. R. (1980). The cognition of discovery: Defining a rhetorical problem. *College Composition and Communication*, 31(1), 21-32.

Graham, S. & Hebert, M. (2011). Writing to read: A meta-analysis of the impact of writing and writing instruction on reading. *Harvard Educational Review*, 81(4), 710-744.

Graham, S., McKeown, D., Kiuhara, S. & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology*, 104(4), 879-896.

Graham, S. & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, 99(3), 445-476.

Graesser, A. C., McNamara, D. S., Louwerse, M. M. & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments & Computers*, 36(2), 193-202.

Grice, H. P. (1975). Logic and conversation. In: P. Cole & J. Morgan (Hrsg.), *Syntax and Semantics: Speech Acts*, 3, 41-58. New York: University Press.

Hachmeister, S. (2019). Messung von Textqualität in Ereignisberichten. In: I. Kaplan & I. Peters. *Schreibkompetenzen messen, beurteilen und fördern*, 79-98. Münster, New York: Waxmann.

Harabagiu, S. M. & Moldovan, D. I. (2000). Enriching the wordnet taxonomy with contextual knowledge acquired from text. In: S. C. Shapiro & L. M. Iwánska (Hrsg.), *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*, 301-333. California: Aaai Press.

Harsch, C., Neumann, A., Lehmann, R. & Schröder, K. (2007). Schreibfähigkeiten. In: B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung: DESI-Studie*, 38-58. Weinheim: Beltz.

Hartig, J. & Frey, A. (2013). Sind Modelle der Item-Response-Theorie (IRT) das „Mittel der Wahl" für die Modellierung von Kompetenzen? *Zeitschrift für Erziehungswissenschaft*, 16(1), 47-51.

Hasselhorn, M., Schumann-Hengsteler, R., Gronauer, J., Grube, D., Mähler, C., Schmid, I., Seitz-Stein, K. & Zoelch, C. (2012). *AGTB 5-12. Arbeitsgedächtnisbatterie für Kinder von 5 bis 12.* Göttingen: Hogrefe.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, 29, 369-388.

Hayes, J. R. & Flower, L. S. (1980). Identifying the organization of writing processes. In: L. W. Gregg & E. R. Steinberg (Hrsg.). *Cognitive Processes in Writing*, 3-30. Hillsdale, NJ: Lawrence.

Heinemann, W. (2000). Textsorte – Textmuster – Texttyp. Text- und Gesprächslinguistik. In: K. Brinker, G. Antos, W. Heinemann & S. Sager (Hrsg.). *Text- und Gesprächslinguistik: Ein internationales Handbuch zeitgenössischer Forschung*, 507-523. Berlin, New York: de Guyter.

Hennes, A.-K., Büyüknarci, Ö., Rietz, C. & Grünke, M. (2015). Helping children with specific learning disability to improve their narrative writing competence by teaching them to use the story maps strategy. *Insights on Learning Disabilities*, 12(1), 35-56.

Hennes, A.-K., Schmidt, B. M., Schabmann, A. & Rietz, C. (2017). Teacher's Expertise in Assessing Written Composition. Spoken paper presented at the Twenty-Fourth Annual Meeting of the Society for the Scientific Study of Reading, Halifax, Canada July, 12-15, 2017.

Hennes, A-K., Schmidt, B. M., Zepnik, S., Linnemann, M., Jost, J., Becker-Mrotzek, M., Rietz, C. & Schabmann, A. (2018). Schreibkompetenz diagnostizieren: Ein standardisiertes Testverfahren für die Klassenstufen 4-9 in der Entwicklung. *Empirische Sonderpädagogik*, 3, 294-310.

Hesse, I. & Latzko, B. (2017). *Diagnostik für Lehrkräfte* (Bd. 3). Opladen: Verlag Barbara Budrich.

He, T.-H., Gou, W. J., Chien, Y.-C., Chen, I.-S. J. & Chang, S.-M. (2013). Multi-faceted Rasch measurement and bias patterns in EFL writing performance assessment. *Psychological Reports,* 112(2), 469-485.

Hodges, T. S., Wright, K. L., Wind, S. A., Matthews, S. D., Zimmer, W. K. & McTigue, E. (2019). Developing and examining validity evidence for the Writing Rubric to Inform Teacher Educators (WRITE). *Assessing Writing*, 40, 1-13.

Holle, K. (2010). Psychologische Lesemodelle und ihre lesedidaktischen Implikationen. In: C. Garbe, K. Holle & T. Jesch (Hrsg.), *Texte lesen. Textverstehen, Lesedidaktik, Lesesozialisation*, 103-165. Paderborn: Ferdinand Schöningh Verlag.

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der Pädagogischen Diagnostik* (Bd. 6). Weinheim: Beltz Verlag.

Jakobs, E. M. & Perrin, D. (2014). *Handbook of Writing and Text Production* (Bd. 10). Berlin; New York: de Gruyter.

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science*, 4(1), 71-115.

Jones, I. & Karadeniz, I., 2016. An alternative approach to assessing achievement. *Proceedings of the 2016 40th Conference of the International Group for the Psychology of Mathematics Education, Szeged, Hungary, 3-7 August 2016.* Prag: International Group for the Psychology of Mathematics Education

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130-144.

Jost, J. & Böttcher, I. (2012). Leistungen messen, bewerten und beurteilen. In: M. Becker-Mrotzek, I. Böttcher & J. Dreher (Hrsg.), *Schreibkompetenz entwickeln und beurteilen*, 113-144. Berlin: Cornelsen.

Kaufman, J. C., Baer, J., Cole, C. J. & Sexton, J. D. (2008). A comparison of expert and non-expert raters using the consensual assessment technique. *Creative Research Journal*, 20(2), 171-178.

Kaufman, J. C., Baer, J. & Cole, J. C. (2009). Expertise, domains, and the consensual assessment technique. *The Journal of Creative Behavior*, 43(4), 223-233.

Kaufman, J. C., Baer, J., Cropley, D. H., Reiter-Palmon, R. & Sinnett, S. (2013). Furious activity vs. understanding: How much expertise is needed to evaluate creative work? *Psychology of Aesthetics, Creativity, and the Arts*, 7(4), 322-340.

Kellogg, R. T. (1996). A model of working memory in writing. In: C. M. Levy & S. E. Ransdell (Hrsg.), *The Science of Writing: Theories, Methods, Individual Differences and Applications*, 57-71. Mahwah, NJ: Erlbaum.

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, 1(1), 1-26.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2), 163-182.

KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland) (2003). *Bildungsstandards im Fach Deutsch für den Mittleren Schulabschluss. Beschluss vom 4.12.2003*. München: Luchterhand.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, 16(2), 81-96.

Knopp, M., Jost, J., Linnemann, M. & Becker-Mrotzek, M. (2014). Textprozeduren als Indikatoren von Schreibkompetenz: Ein empirischer Zugriff. In: T. Bachmann & H. Feilke (Hrsg.), *Werkzeuge des Schreibens: Beiträge zu einer Didaktik der Textprozeduren*, 111-128. Stuttgart: Klett.

Knopp, M., Jost, J., Nachtwei, N., Becker-Mrotzek, M. & Grabowski, M. (2012). Teilkomponenten von Schreibkompetenz untersuchen: Bericht aus einem interdisziplinären empirischen Projekt. In: H. Bayrhuber, U. Harms & B. Muszynski (Hrsg.), *Formate Fachdidaktischer Forschung: Empirische Projekte – historische Analysen – theoretische Grundlegungen*, 47-65. Münster: Waxmann.

Krifka, M. & Musan, R. (2012). Information structure: Overview and linguistic issues. In: M. Krifka & R. Musan (Hrsg.), *The Expression of Information Structure*. Berlin, New York: de Gruyter.

Koster, M., Tribushinina, E., De Jong, P. F. & Van den Bergh, H. (2015). Teaching children to write: A meta-analysis of writing intervention research. *Journal of Writing Research*, 7(2), 249-274.

Leckie, G. & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48(4), 399-418.

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V. & De Maeyer, S. (2017). Comparative judgement as a promising alternative to score competences. In: E. Cano & G. Ion (Hrsg.) *Innovative Practices for Higher Education Assessment and Measurement*, 119-138. Hershey, PA: IGI Global.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16(2), 878.

Linacre, J. M. (2003). Size vs. significance: Standardized Chi-Square fit statistic. *Rasch Measurement Transactions*, 17(1), 918.

Linacre, J. M. (2012). Many-facet Rasch measurement: Facets tutorial. (https://www.winsteps.com/a/ftutorial1.pdf; zuletzt abgerufen am 19.03.2019).

Linacre, J. M. (2019). *A User's Guide to FACETS: Rasch-Model Computer Programs*. Chicago: Winsteps.com.

Linacre, J. M. & Wright, B. D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions*, 8(2), 350.

Linke, A. & Nussbaumer, M. (2000). Konzepte des Impliziten: Präsuppositionen und Implikaturen. In: K. Brinker, G. Antos, W. Heinemann & S. F. Sager (Hrsg.). *Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung*, Bd. 1, 435-448. Berlin, New York: de Gruyter.

Linnemann, M. & Stephany, S. (2014). Supportive writing assignments for less skilled writers in mathematic classroom. In: P. Klein, L. Boscolo, S. Kirkpatrick & C. Gelati (Hrsg.), *Studies in Writing: Writing as a Learning Activity*, 6-93. Leiden: Koninklijke Brill NV.

Luginbühl, M. & Perrin, D. (2011). Muster und Variation. In: M. Luginbühl & D. Perrin (Hrsg.), *Medienlinguistische Perspektiven auf Textproduktion und Text.* Frankfurt am Main: Peter Lang.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, 19(3), 246-276.

MacArthur, C. A., Graham, S. & Fitzgerald, J. (2016). *Handbook of Writing Research*. New York: Guilford Press.

MacArthur, C. A. & Graham, S. (2016). Writing research from a cognitive perspective. In: C. A. MacArthur, S. Graham & J. Fitzgerald (Hrsg.), *Handbook of Writing Research*, 24-40. New York: Guilford Press.

MacArthur, C. A., Jennings, A. & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing*, 32(6), 1553-1574.

Marsh, H. W. (1987) The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, 79(3), 280-295.

McNamara, D. S., Kintsch, E., Butler-Songer, N. & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14, 1-43.

McNamara, T., Knoch, U. & Fan, J. (2019). *Fairness, Justice & Language Assessment*. Oxford: University Press.

McNamara, D.S., Louwerse, M. M., McCarthy, P. M. & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47, 292-330.

Murphy, S. & Yancey, K. B. (2008). Construct and consequence: Validity in writing assessment. In: C. Bazermann (Hrsg.), *Handbook of Research on Writing: History, Society, School, Individual, Text*, 365-385. New York: Taylor & Francis.

NAEP (2011): National Assessment Governing Board (2011). *Writing Framework for the 2011 National Assessment of Educational Progress.* U.S. Department of Education, Washington, D.C.

National Commission on Writing. (2003). *The Neglected "r": The Need for a Writing Revolution.* New York: College Entrance Examination Board.

Nussbaumer, M. (1991). *Was Texte sind und wie sie sein sollen. Ansätze zu einer sprachwissenschaftlichen Begründung eines Kriterienrasters zur Beurteilung von schriftlichen Schülertexten.* Tübingen: Niemeyer.

Nussbaumer, M. & Sieber, P. (1995): Über Textqualitäten reden lernen – z.B. anhand des „Zürcher Textanalyserasters". *Diskussion Deutsch*, 26, 36-52.

Olive, T., Kellogg, R. T. & Piolat, A. (2008). Verbal, visual, and spatial working memory demands during text composition. *Applied Psycholinguistics*, 29, 669-687.

Ohlinghouse, N. G., Santangelo, T. & Wilson, J. (2012). Examining the validity of single-occasion, single-genre, holistically scored writing assessments. In: E. van Steendam, M. Tillema, G. Rijlaarsdam & H. van den Bergh (Hrsg.), *Measuring Writing. Recent Insights into Theory, Methodology and Practice*, 55-82. Leiden: Brill.

O'Reilly, T. & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes*, 43, 121-152.

Orell Füssli Sicherheitsdruck AG (2010). *Papa Moll, Band 10, PM die Sportskanone*. Zürich: Globi Verlag.

Ossner, Jakob (2006) Kompetenzen und Kompetenzmodelle im Deutschunterricht. *Didaktik Deutsch*, 21, 5-19.

Penner-Williams, J., Smith, T. E. & Gartin, B. C. (2009). Written language expression: Assessment instruments and teacher tools. *Assessment for Effective Intervention*, 34(3), 162-169.

Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R. & Richardson, C. (1995). A redrawn Vandenberg & Kuse Mental Rotations Test: Different versions and factors that affect performance. *Brain and Cognition*, 28, 39-58.

Piaget, J. (1973). *The Child and Reality*. New York: Viking Press.

Philipp, M. (2015). *Schreibkompetenz: Komponenten, Sozialisation und Förderung*. Tübingen: A. Francke Verlag.

Pohl, T. (2007). Emotionalität im frühen Schreiben – Von emotionaler Involviertheit zu emotionaler Involvierung. In: M. Becker-Mrotzek & K. Schindler (Hrsg.). *Texte schreiben*. *KöBeS* (Bd. 5), 63-80. Duisburg: Gilles & Francke Verlag.

Pohl, T. & Steinhoff, T. (2010). *Textformen als Lernformen*. Duisburg: Gilles & Francke.

Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22(2), 157-170.

Primi, R., Silvia, P. J., Jauk, E. & Benedek, M. (2019). Applying many-facet Rasch modeling in the assessment of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2), 176-186.

Rezaei, A. R. & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18-39.

Rickheit, G. & Strohner, H. (2003). Inferenzen. In: G. Rickheit, T. Herrmann & W. Deutsch (Hrsg.), *Psycholinguistik – Psycholinguistics. Ein internationales Handbuch*, 566-577. Berlin, New York: de Gruyter.

Sakyi, A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In: A. J. Kunnan (Hrsg.), *Studies in Language Testing. Fairness and Validation in Language Assessment*, 129-152. Cambridge: University Press.

Sandig, B. (1997). Formulieren und Textmuster. Am Beispiel von Wissenschaftstexten. In: E. M. Jakobs, & D. Knorr (Hrsg.), *Schreiben in den Wissenschaften,* 25-44. Frankfurt am Main: Peter Lang.

Scardamalia, M. & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. *Advances in Applied Psycholinguistics*, 2, 142-175.

Seker, M. (2018). Intervention in teachers' differential scoring judgments in assessing L2 writing through communities of assessment practice. *Studies in Educational Evaluation*, 59, 209-217.

Schindler, K. (2004). *Adressatenorientierung beim Schreiben. Eine linguistische Untersuchung am Beispiel des Verfassens von Spielanleitungen, Bewerbungsbriefen und Absagebriefen.* Frankfurt am Main: Peter Lang.

Schmitt, M. (2011). *Perspektivisches Denken als Voraussetzung für adressatenorientiertes Schreiben.* Dissertation, Pädagogische Hochschule Heidelberg. (http://nbn-resolving.de/urn:nbn:de:bsz:he76-opus-75266; zuletzt abgerufen am 29.12.2019).

Schmitt, M. & Knopp, M. (2017). Prädiktoren der Schreibkompetenz. In: M. Becker-Mrotzek, J. Grabowski & T. Steinhoff (Hrsg.), *Forschungshandbuch empirische Schreibdidaktik*, 239-252. Münster, New York: Waxmann.

Schneider, W., Schlagmüller, M. & Ennemoser, M. (2017). *Lesegeschwindigkeits- und -verständnistest für die Klassenstufen 5-13 (LGVT 5-12+).* Göttingen: Hogrefe.

Schwarz, M. (2001). Kohärenz – Auf den materiellen Spuren eines mentalen Phänomens. In: M. Bräunlich, B. Neuber & B. Rues (Hrsg.), *Gesprochene Sprache – transdisziplinär. Festschrift für Gottfried Meinhold*, 151-160. Frankfurt am Main: Peter Lang.

Schwarz-Friesel, M. (2006). Kohärenz versus Textsinn: Didaktische Facetten einer linguistischen Theorie der textuellen Kontinuität. In: M. Scherner & A. Ziegler (Hrsg.), *Angewandte Textlinguistik. Perspektiven für den Deutsch- und Fremdsprachenunterricht, Europäische Studien zur Textlinguistik*, Bd. 2, 63-75. Tübingen: Gunter Narr.

Shanahan, T. (2006). Relations among oral language, reading, and writing development. In: C. A. MacArthur, S. Graham & J. Fitzgerald (Hrsg.), *Handbook of Writing Research*, 171-183. New York, London: Guilford Press.

Shanahan, T., Callison, K., Carriere, C., Duke, N. K., Pearson, P. D., Schatschneider, C. & Torgesen, J. (2010). *Improving Reading Comprehension in Kindergarten through 3rd Grade: A Practice Guide (NCEE 2010-4038).* Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Smith, M., Cheville, J. & Hillocks, G. (2006). I guess I'd better watch my English. In: C. A. MacArthur, S. Graham & J. Fitzgerald (Hrsg.), *Handbook of Writing Research*, 263-274. New York, London: Guilford Press.

Sripada, S. G., Reiter, E., Hunter, J. & Yu, J. (2003). Generating English summaries of time series data using the Gricean maxims. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 187-196. New York: ACM Press.

Steinert, J. (2011). *Allgemeiner Deutscher Sprachtest (ADST)*. Göttingen: Hogrefe.

Strong, J. Z. (2019). *Effects of a Text Structure Intervention for Reading and Writing in Grades 4-5: A Mixed Methods Experiment*. Doctoral dissertation, University of Delaware. (https://search.proquest.com/docview/2307786372?accountid=10218; zuletzt abgerufen am 17.02.2020).

Sternberg, R. J. (1999). *Handbook of Creativity*. Cambridge: University Press.

Sternberg, R. J., Kaufman, J. C. & Pretz, J. E. (2002). *The Creativity Conundrum*. New York: Psychology Press.

Sturm, A. & Weder, M. (2016). *Schreibkompetenz, Schreibmotivation, Schreibförderung. Grundlagen und Modelle zum Schreiben als soziale Praxis.* Seelze: Kallmeier.

Thonke, F., Groß Ophoff, J., Hosenfeld, I. & Isaac, K. (2008). Kriteriengestützte Erfassung von Schreibleistungen im Projekt VERA. *Checkpoint Literacy. Tagungsband 2 zum 15. Europäischen Lesekongress in Berlin,* Bd. 2, 28-35. Berlin: Deutsche Gesellschaft für Lesen und Schreiben.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273-286.

Todd, R. W., Thienpermpool, P. & Keyuravong, S. (2004). Measuring the coherence of writing using topic-based analysis. *Assessing writing*, 9(2), 85-104.

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O. & Baumert, J. (2006). Tracking, grading and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, 98(4), 788-806.

Twain, M. (2009). *Huckleberry Finn*. Frankfurt am Main, Leipzig: Insel Taschenbuch.

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2016). Validity of comparative judgement to assess academic writing: Examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice*, 26 (1), 59-74.

Van Dijk, T. A. (1995). On macrostructures, mental models, and other inventions: A brief personal history of the Kintsch-van Dijk theory. In: C. A. Weaver, S. Mannes & C. Fletcher (Hrsg.), *Discourse Comprehension: Essays in Honor of Walter Kintsch*, 383-410. Hillsdale, NJ: Lawrence Erlbaum.

Van Dijk, T. A. & Kintsch, W. (1983). *Strategies of Discourse Comprehension*. New York: Academic Press.

Vögelin, C., Jansen, T., Keller, S. D. & Möller, J. (2018). The impact of vocabulary and spelling on judgments of ESL essays: An analysis of teacher comments. *The Language Learning Journal*, 1-17.

Vögelin, C., Jansen, T., Keller, S. D., Machts, N. & Möller, J. (2019). The influence of lexical features on teacher judgements of ESL argumentative essays. *Assessing Writing*, 39, 50-63.

Voss, J. F., Fincher-Kiefer, R. H., Greene, T. R. & Post, T. A. (1986). Individual differences in performance: The contrastive approach to knowledge. *Advances in the Psychology of Human Intelligence*, 3, 297-334.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: University Press.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11, 197-223.

Weiß, R. H. (2006). *Grundintelligenztest Skala 2-revision (CFT 20-R)*. Göttingen: Hogrefe.

Whitehouse, C. & Pollitt, A. (2012). *Using Adaptive Comparative Judgement to Obtain a Highly Reliable Rank Order in Summative Assessment*. Manchester: AQA Centre for Education Research and Policy.

Wimmer, H. & Mayringer, H. (2014). *Salzburger Lese-Screening für die 2.-9. Klasse* (SLS 2-9). Bern: Huber.

Wirtz, M. A. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

Wright, B. D. & Masters, G. N. (2002). Number of person or item strata. *Rasch Measurement Transactions*, 16(3), 888.

Wu, M. (2017). Some IRT-based analyses for interpreting rater effects. *Psychological Test and Assessment Modeling*, 59(4), 453-470.