

Christopher Sappok

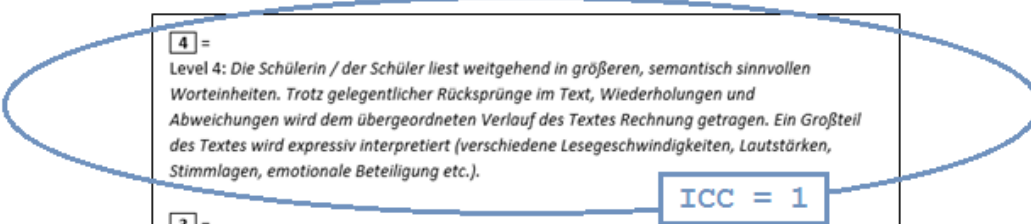
Exploring Advanced Prosody – eine Best-Practice- Untersuchung zum lauten Lesen in der weiterführenden Schule

1 | Einführung

Der vorliegende Beitrag stellt eine Ratingprozedur mit drei Skalen zur Erhebung von fortgeschrittener Vorlesekompetenz und ihren prosodischen Komponenten vor. Anhand von 156 kurzen Aufnahmeausschnitten wurden „Vorlesewettbewerb“ (Skala C) und als hypothetische Komponenten von C die Faktoren „Stimmführung“ (Skala B) und „Hervorhebung“ (Skala A) erhoben. Dies geschah zum einen, um die Skalen zu evaluieren und zum anderen, um unseren Begriff von voll ausgebildeter Leseflüssigkeit zu konkretisieren¹. Generell gilt nach aktuellem Forschungsstand nach wie vor das *Henne-Ei-Dilemma* (Nix 2011; Kuhn, Schwanenflugel & Meisinger 2010): Hilft *reading prosody* beim Textverstehen oder setzt *reading prosody* ein Textverstehen voraus? Damit stellt sich die Frage, welchen über sich selbst hinausweisenden, also auf Textverstehen bezogenen diagnostischen Wert die Ermittlung und welchen didaktischen Wert die Vermittlung von Vorlese-Skills hat (Funke 2018).

Spätestens mit Holle (2006) wird die Relevanz von *reading aloud* oder *oral reading* (Daane et al. 2005) auch im deutschsprachigen Raum immer höher gewertet (Rosebrock & Nix 2006; Rosebrock et al. 2016). Untersucht werden in diesen Kontexten hauptsächlich Audiodaten lesender Grundschüler*innen (Sappok & Fay 2018; Röttig et al. und Stephany et al. in diesem Band). Längsschnittstudien (Röttig et al. in diesem Band) sind die Ausnahme und für die Zeit nach der Grundschule liegen kaum empirische Untersuchungen vor, auch international (Godde, Bosse & Bailly 2020). Im vorliegenden Beitrag werden Schüler*innen untersucht, die in einer früheren Untersuchung (Sappok, Linnemann & Stephany 2020) als besonders flüssig identifiziert wurden (Level 4 auf der Pinnell-Skala, s. Abbildung 1). Die Untersuchung einer Best-practice-Stichprobe soll Aufschluss darüber geben, was flüssige Leser*innen bereits können und worin sie sich dabei immer noch unterscheiden. Dies geschieht auch in Hinblick auf die Operationalisierung in Form von signalphonetischen und damit in Zukunft auch automatisch analysierbaren Parametern. Vor diesem Hintergrund sollen zunächst einige Vorüberlegungen angestellt werden.

¹ Die Audiodaten können für wissenschaftliche Zwecke beim Autor zum Download angefordert werden.



4 = Level 4: Die Schülerin / der Schüler liest weitgehend in größeren, semantisch sinnvollen Worteinheiten. Trotz gelegentlicher Rücksprünge im Text, Wiederholungen und Abweichungen wird dem übergeordneten Verlauf des Textes Rechnung getragen. Ein Großteil des Textes wird expressiv interpretiert (verschiedene Lesegeschwindigkeiten, Lautstärken, Stimmlagen, emotionale Beteiligung etc.).	ICC = 1
3 = Level 3: Die Schülerin / der Schüler liest überwiegend in Dreier- oder Vierer-Wortgruppen; gelegentlich treten auch kleinere Wortgruppen auf. Die Mehrheit der Wortgruppierung ist (trotzdem) angemessen und entspricht der Syntax des Textes. Wenig oder keine expressive Interpretation (Verschiedene Lesegeschwindigkeiten, Lautstärken, Stimmlagen, emotionale Beteiligung etc.).	
2 = Level 2: Die Schülerin / der Schüler liest überwiegend in Zweier-Wortgruppen. Dreier- und Vierer-Wortgruppen treten gelegentlich auf. Ab und zu kommt auch ein Wort-für-Wort Lesen vor. Die Wortgruppierungen erscheinen ungeschickt und stehen in keinem Zusammenhang zur Syntax des Textes.	
1 = Level 1: Die Schülerin / der Schüler liest den Text hauptsächlich Wort für Wort. Nur gelegentlich treten Zweier- oder Dreier-Wortgruppierungen auf. Die wenigen Wortgruppierungen sind unregelmäßig und unterstützen nicht die Syntax des Textes.	

Abb. 1: Die NAEP-Leseflüssigkeitsskala in adaptierter Form (Pinnell et al. 1995; Rosebrock & Nix 2015). Drei von drei Rater*innen haben die unten untersuchten Schüler*innen mit NAEP-Fluency = 4 bewertet (Sappok, Linnemann & Stephany 2020), was einem Reliabilitätswert von (post hoc) ICC = 1 entspricht.

Der Titelbegriff *advanced prosody* setzt einen *early prosody*-Begriff voraus. Für den deutschsprachigen Raum kann die prosodische Dimension des Konstrukts Leseflüssigkeit mit *early prosody* insoweit gleichgesetzt werden, als dass Leseflüssigkeit überwiegend im Primarkontext bzw. im Kontext frühen Textverstehens fokussiert wird. Neuere Erkenntnisse sprechen jedoch dafür, hiervon *advanced prosody* dezidiert zu unterscheiden.

Einen weitgehend anerkannten Ausgangspunkt stellt die Fluency-Definition mit drei Dimensionen von Kuhn, Schwanenflugel und Meisinger (2010, Herv.: CS) dar:

„Fluency combines *accuracy, automaticity, and oral reading prosody*, which, taken together, facilitate the reader’s construction of meaning. It is demonstrated during oral reading through ease of word recognition, appropriate pacing, phrasing, and intonation. It is a factor in both oral and silent reading that can limit or support comprehension.“

Groen, Veenendaal & Verhoeven (2018) kritisieren an dieser Definition, dass *automaticity* und *reading prosody* ein gleichrangiger Stellenwert in Bezug auf *reading comprehension* zugewiesen wird. Sie verglichen eine Gruppe niederländischer, älterer *poor comprehenders* (Jahrgangsstufe 5), die dabei altersgemäß automatisiert lasen, mit einer gleichaltrigen Gruppe, bei der auch *comprehension* altersgemäß war. Die Autor*innen stellten fest, dass bei den *poor comprehenders* auch der Faktor *text reading prosody*, der sehr differenziert erhoben wurde, „poor“ ausgeprägt ist, und fassen zusammen:

„It has been proposed that reading fluency – as a combination of accuracy, automaticity and text reading prosody – facilitates the reader’s construction of meaning (Kuhn et al. 2010). The results from the current study [...] suggest that the ‘automaticity aspect’ of reading is a distinct process from the construction of meaning. The construction of meaning seems more closely tied to text reading prosody than to decoding efficiency, at least, when children have mastered automaticity in reading” (Groen, Veenendaal & Verhoeven 2018: 16).

Die zuletzt genannte Bedingung kann für die Stimmen, die auf den Aufnahmen der vorliegenden Untersuchung zu hören sind, als voll erfüllt gelten. So können prosodische Merkmale in den Blick genommen werden, ohne dass ihre Wahrnehmung durch andere Defizite gestört wird. Zur Vertiefung soll zunächst eine differenzierte Klärung zum allgemeinen Prosodiebegriff vorgenommen werden.

Um Prosodie und deren Verhältnis zu linguistischen Einheiten zu verdeutlichen, ist die Unterscheidung von vier *perceptual domains* von Laver (1994) die beste Grundlage: „There [are] only four perceptual domains available to the human auditory system for differentiating the elements of speech. These [are] the domains of perceptual quality, duration, pitch and loudness” (Laver 1994: 431).

Quality steht hier für das, *was* gesagt wird und damit für den linguistischen Gehalt, wie ihn der Wortlaut des Textstimulus liefert. Prosodie hingegen steht als Sammelbegriff für die prosodischen Basismerkmale *duration* (Dauerphänomene), *pitch* (Tonhöhenbewegung) und *loudness* (Lautheitsphänomene), kurz: dafür, *wie* etwas gesagt wird.

Abbildung 2 stellt die drei prosodischen Basismerkmale als nominal skalierte Z-Achse in erweiterte Zusammenhänge.

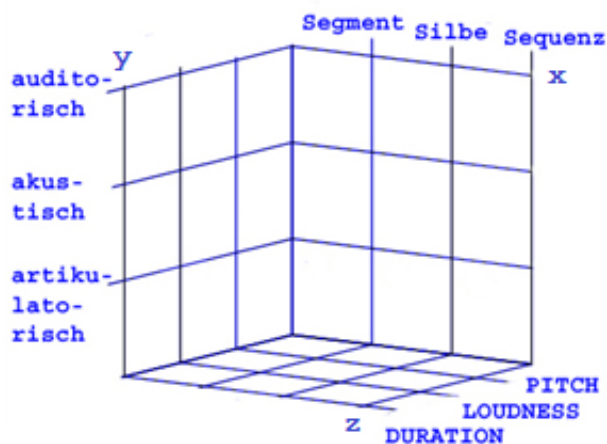


Abb. 2: Prosodie als dreidimensionales Konstrukt. Die X-Achse bezeichnet Größenordnungen (Tillmann & Mansell 1980), die Y-Achse bezeichnet Korrelate (Pompino-Marschall 2009) und die Z-Achse bezeichnet Basismerkmale (Laver 1994) von Prosodie. Es ergibt sich ein Raum zur Verortung von Bestimmungsstücken prosodischer Forschung.

Die X-Achse Größenordnung unterscheidet nicht nur die Bezugseinheiten Segment, Silbe und Sequenz, sondern auch unterschiedliche Zeitfenster der kognitiven Sprachverarbeitung (Miller 1956; Kegel 1990; Pöppel 1997; Sappok 2010). Geht man hier hinsichtlich der Fensterbreite von

Segment = 30 ms, Silbe = 300 ms und Sequenz = 3000 ms aus, kann die X-Achse als logarithmisch skaliert gelten. Sie geht auf die Prosodiekonzeption von Tillmann zurück (Tillmann & Mansell 1980).

Hinter der Y-Achse stehen drei Typen von phonetischen Korrelaten. Diese Unterscheidung leitet sich aus der nachrichtentechnischen Unterscheidung von Sender, Kanal und Empfänger her (Weaver & Shannon 1949). Auch wenn das Kommunikationskettenmodell eine zeitliche Abfolge suggeriert, ist die Y-Achse als nominal skaliert zu bezeichnen, weil die Einteilung vor allem Methodenunterschiede bei der artikulatorischen, akustischen und auditiven Phonetik (Pompino-Marschall 2009) in den Vordergrund rückt, die in keinem quantifizierbaren Verhältnis zueinander stehen.

Damit ergeben sich 27 hypothetische prosodische Einzelmerkmale als Punkte im XYZ-Raum. Sie sind nicht alle gleichermaßen relevant, bieten aber einen differenzierten Rahmen zur Verortung von Bestimmungstücken prosodischer Forschung allgemein. Gravierend zu nennen ist die Einschränkung, dass pitch, loudness und duration hier in allen Kontexten gleichermaßen isoliert dastehen. Mit aufsteigender Größenordnung gilt jedoch mehr und mehr, dass die drei Basismerkmale zusammenwirken. Dennoch wird pitch in der internationalen Prosodieforschung eine besonders große Rolle zugeschrieben und auch in Hinblick auf advanced prosody kann pitch als besonders aufschlussreich gelten.

In der Grundlagenforschung zur Schnittstelle von Prosodie und Sprachverstehen steht die Untersuchung explizit vorgelesener Sprache nicht im Vordergrund (eine Ausnahme ist Blaauw 1995), sondern Sprache im Diskurs. Einen aktuellen Forschungsüberblick hierzu bietet Dahan (2015). Dort heißt es:

“Prosody is characterized as an abstract structure composed of discrete tonal elements aligned with the segmental composition of the sentence organized in constituents of increasing size, and this structure is influenced by the phonological, syntactic, and informational structures of the sentence” (Dahan 2015: 441).

Hervorzuheben ist hier, dass *informational structure* (oder Fokus, vgl. Fisseni 2011) besonders nah mit Textverstehen und einer „singgestaltenden“ oder gar „sinnstiftenden“ Leseprosodie in Verbindung gebracht werden kann.

Ein Modell zu pitch, das im Kontext der Sprachsyntheseforschung besonders einflussreich ist, ist das Fujisaki-Modell (Fujisaki 1988; für das Deutsche: Möbius 1993, s. Abbildung 3), an dem sich die vorliegende Untersuchung orientiert. Im Fujisaki-Modell wird pitch als das additive Zusammenwirken von zwei Steuermechanismen, *phrase component* (Sequenz) und *accent component* (Silbe), modelliert. Addiert ergeben die Outputs dieser Mechanismen Grundfrequenz, also einen akustischen Parameter (F0). Die erste Komponente liefert Eckparameter für eine globale Tendenz (T01 und T02 in Abbildung 3), die meist abfallend ist („Deklination“, Pompino-Marschall 2009). Die zweite liefert Eckparameter für diskrete Auslenkungsintervalle (T11 bis T32 in Abbildung 3). Das additive Zusammenspiel der Komponenten wird Superimposition genannt (Fujisaki 1988).

Im vorliegenden Kontext ist für die Accent-Komponente die Skala A zuständig, beschränkt sich aber im Gegensatz zum Fujisaki-Modell auf einen Hauptakzent. Skala B nimmt dem Modell entsprechend die Phrasenkomponente in den Blick.

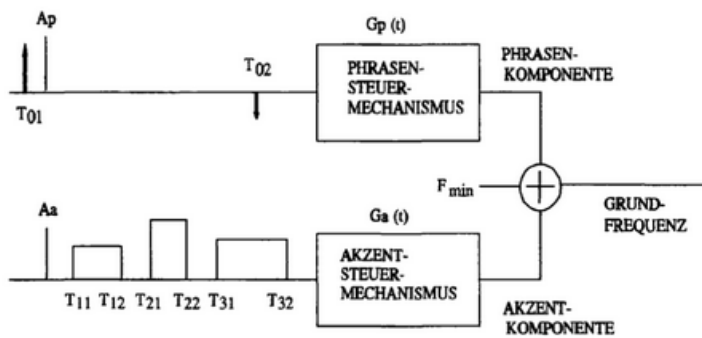


Abb. 3: Fujisaki-Modell mit zwei additiven Komponenten von Grundfrequenz (Fujisaki 1988, Reproduktion aus Möbius 1993: 68).

Geht es nun speziell um das laute Lesen in didaktischen Kontexten, spielen der Text und seine spezifischen prosodischen Herausforderungen eine große Rolle. In Sappok & Fay (2018) wurde mit dem einfacheren narrativen Text „Der Hase und der Schneemann“ mit 12 Drittklässler*innen gearbeitet. Erhoben und über signalphonetische Messungen rekonstruiert wurde, wie diese Schüler*innen Sprecherwechsel in reinem Dialog gestalten (Diskursgliederung). Hierfür wurde jede Gesamtaufnahme in ca. 2 bis 4 Sekunden lange Doppelsequenzen geschnitten. Zu den Doppelsequenzen wurden Ratings durchgeführt und signalphonetisch mit Bezug zu pitch, loudness und duration akustische Korrelate von Diskursgliederung gemessen (Stimme-Verstellen und Pausen). Ein wesentlicher Befund bestand darin, dass auch innerhalb von Pinnell-Level 4 ein breites Leistungsspektrum aufgezeigt werden konnte. Unter ähnlichen methodischen Bedingungen wurde die hier vorgestellte Ratingprozedur organisiert. Hierauf wird nun kurz eingegangen:

Der Text „Dolmetscher“ (s. drittes Unterkapitel), der im Zentrum der vorliegenden Untersuchung steht, kann im Gegensatz zu „Der Hase und der Schneemann“ (Sappok & Fay 2018) als besonders schwierig gelten. Statt prosodischer Diskursgliederung steht mit Skala C eine maximal holistische Vorlesekompetenz als nicht weiter in Frage gestellte abhängige Variable im Zentrum. Vorlesekompetenz ist hier also genau das, was die Vorlesekompetenz-Skala misst.

Vor diesem Hintergrund lässt sich für die vorliegende Untersuchung zusammenfassen:

- zentrale Hypothese: Die erhobenen Ratingdaten können zu einem explorativen Modell nach dem Muster $C \sim B + A +$ weitere Faktoren integriert und methodisch und didaktisch weiterführend interpretiert werden.

Die Differenzierung dieser Komponenten stellt hohe methodische Anforderungen. Dies gilt besonders für Faktor A (Sappok 2002; Sappok 2010; Sappok & Arnold 2012a, 2012b).

Aus dieser Grundlagen-Sicht heraus bzw. im Hinblick auf eine Evaluation der Skalen werden neben den Lautlese-Performanzen von Schüler*innen auch die Hör-Performanzen von Rater*innen (N = 39 Lehramtstudierende) genauer in den Blick genommen. Im Hintergrund stehen dabei zwei Annahmen, die als Heurismen wie folgt formuliert werden können:

- Heurismus I: Das besondere Problem bei der diagnostischen Analyse von Audiodaten besteht eher in einem Überangebot an potenziell relevanter Information als in einem Defizit.

- Heurismus II: Vorlesekompetenz und ihre Komponenten lassen sich effektiver auf Phrasen-Ebene anstatt auf Ebene des gesamten Textes erfassen.

Die Motivation des Schneidens *in* bzw. Beurteilen-Lassens *von* 2-4 Sekunden Sequenzen besteht darin, dass damit Entscheidungen stimuliert werden, die direkt aus dem Arbeitsgedächtnis abgerufen werden können (entspricht einem Fenster von etwa 3 Sekunden „unmittelbarer Gegenwart“, Pöppel 1997). Das Rating beschränkt sich also auf die Sequenzebene und erspart es dem Hörenden, über den ganzen Text zu integrieren (eine Textebene gibt es wohlweislich nicht in Abbildung 2). Erwartet wird so eine höhere Differenziertheit der Ratings.

Aus weiterführend-didaktischer Sicht stehen hier wie bei Sappok & Fay (2018) im Hintergrund zwei Annahmen, die als spekulative Axiome wie folgt formuliert werden können:

- Axiom I: Kompetenz beim lauten Lesen ist ein Indikator für Aspekte von Textverstehen.
- Axiom II: Umgekehrt können solche Aspekte über die Förderung von lautem Lesen mitgefördert werden („bootstrapping“).

In vorliegender Untersuchung steht die Diagnostik im Vordergrund, über die Axiome kann nur weiterführend spekuliert werden. Dies wird unter Einbeziehung der didaktischen Perspektive im fünften und letzten Unterkapitel aufgegriffen. Zunächst (Unterkapitel 2) werden Stichprobe und Aufnahmebedingungen näher beschrieben. Dazu werden die Anknüpfungspunkte der Einführung zu konkreteren Hypothesen weiterentwickelt. Das dritte Unterkapitel behandelt den Textstimulus und erläutert die drei Ratingskalen sowie deren computergestützten Einsatz bei der Evaluation. Im vierten Unterkapitel werden die dabei erhobenen Ratingdaten analysiert und interpretiert. Im Zentrum des Abschlusskapitels steht, wie gesagt, die Deutschdidaktik.

2 | Stichprobe und Hypothesen

Die 26 Aufnahmen von 16 Schüler*innen, die hier untersucht werden, stammen aus einem Korpus mit der Bezeichnung LAUDIO (= „Longitudinal Audio“), das insgesamt knapp 1000 Aufnahmen von 61 Schüler*innen enthält, die vom Autor dieses Beitrags über einen Zeitraum von mehreren Jahren beim lauten Lesen aufgezeichnet wurden. Dabei kamen verschiedene Textstimuli zum Einsatz, die eine große Bandbreite von relevanten Textmerkmalen repräsentieren (Schwierigkeitsgrad, Länge, Genre usw.). Bei dem Textstimulus „Dolmetscher“ (im Folgenden: DOL), mit dem die hier untersuchten Aufnahmen entstanden sind, handelt es sich um einen kurzen, aber vergleichsweise anspruchsvollen Sachtext, der im Gegensatz zu den anderen verwendeten Texten als Ankertext bei allen Aufnahmesitzungen zum Einsatz kam (außer wenn ad hoc eindeutig entschieden werden konnte, dass dieser Text eine massive Überforderung darstellen würde, v. a. in Jahrgangsstufe 3). Auf diesen Text wird im nächsten Unterkapitel näher eingegangen. Im Folgenden wird zunächst die Stichprobe der gesamten LAUDIO-Datenerhebung und die Auswahl der hier untersuchten Aufnahmen näher beschrieben.

Die LAUDIO-Datenerhebung begann gegen Ende des ersten Halbjahres 2014/2015. Es wurden zwei dritte und zwei vierte Klassenverbände einer inklusiven Grundschule im ländlichen Raum von Rheinland-Pfalz aufgenommen, und zwar immer in Einzelsitzungen mit dem Versuchsleiter (CS) und immer mit der mündlichen Instruktion: „Hier hast du einen Text. Lies mir den Text jetzt bitte mal vor.“ bzw. beim zweiten Lesen – stets unabhängig von der Qualität der eben erfolgten

Performance: „Super gemacht! Lies mir den Text jetzt bitte noch mal vor, da klappt’s dann bestimmt *noch* besser.“

Insgesamt hatten die vier Klassenverbände 73 Schüler*innen, wovon zum ersten Erhebungszeitpunkt 61 Schüler*innen an der Untersuchung teilnahmen. Die Differenz ist auf fehlendes Einverständnis der Eltern oder auf krankheitsbedingtes Fehlen zurückzuführen und wird im Folgenden als zufallsbedingt erachtet. Begleitend zur ersten Audio-Datenerhebung wurden die jeweiligen Deutschlehrkräfte gebeten, zu jedem Kind Einschätzungen in Bezug auf acht Teilkompetenzen des Deutschunterrichts auf der klassischen Notenskala von *sehr gut* bis *ungenügend* abzugeben. Erfragt wurden hier die Aspekte *Leseflüssigkeit*, *Leseverstehen*, *Schreibkompetenz Texte*, *Rechtschreibung*, *Gesprächskompetenz*, *Zuhörkompetenz*, *Sprachreflexionskompetenz* und *Wortschatz* (vgl. KMK 2004, Mittelwerte hieraus erscheinen als Deutschleistung in Tabelle 1). Außerdem wurden Einschätzungen zum Sprachniveau Deutsch, zur Sozialkompetenz (hier nicht einbezogen) und zum sozioökonomischen Hintergrund (hier nicht einbezogen) der Kinder eingeholt. Zusätzlich wurden mit allen jeweils aufgenommenen Kindern zum ersten und allen folgenden Erhebungszeitpunkten schriftliche Tests zur Erfassung verschiedener Aspekte von Lesekompetenz durchgeführt, nämlich ELFE 1 - 6² (Lenhard & Schneider 2006) und SLS 2 – 9 (Wimmer & Mayringer 2014).

In den anschließenden drei Jahren wurden möglichst viele dieser Schüler*innen in durchschnittlich etwa 15-minütigen Einzelsitzungen einmal jährlich erneut aufgenommen, so dass mit dem Ende des Erhebungszeitraums Anfang 2018 Längsschnittdaten zu einer „Kohorte 3 bis 6“ und einer „Kohorte 4 bis 7“ vorliegen. Für eine erste Überblicksuntersuchung (Sappok, Linnemann & Stephany 2020) wurden aus der ursprünglichen Stichprobe insgesamt 31 Schüler*innen ausgewählt ($n_{3-6} = 13$, $n_{4-7} = 18$) und deren DOL-1-Performanzen analysiert, d. h. zum ersten Erhebungszeitpunkt „prima vista“ bzw. „prima vista seit letztem Jahr“ für alle weiteren Erhebungszeitpunkte. Einbezogen wurden in Sappok, Linnemann & Stephany (2020) genau diejenigen Schüler*innen, von denen mindestens zum ersten und zum letzten der vier Erhebungszeitpunkte mindestens eine DOL-Aufnahme vorliegt. Die Ausfälle kommen dadurch zustande, dass zum ersten Erhebungszeitpunkt keine DOL-Aufnahme gemacht wurde (extrem schwache Schüler*innen). Abgesehen davon werden alle anderen Ausfallgründe für die Analyse als zufallsbedingt erachtet (dass einzelne Kinder auf weit entfernte weiterführenden Schulen übergegangen sind, dass das Elterneinverständnis entzogen wurde usw.).

Um diese Zufallsbedingtheit zu überprüfen und damit auch, inwiefern diese Stichprobe trotz der Ausfälle als repräsentativ angesehen werden kann, wurden die ELFE-Testdaten des ersten Erhebungszeitpunkts mit den betreffenden Angaben zur Normierungsstichprobe von ELFE verglichen. Der Vergleich zeigte, dass die Stichprobe keine nennenswerten Abweichungen aufweist, außer dass die Streuung bei den Drittklässler*innen beim Wort-, Satz- und Textverstehen überall leicht höher, bei den Viertklässler*innen hingegen leicht niedriger ausfällt als bei den Normierungen. Insgesamt spricht also wenig gegen die Annahme, dass die Ausprägungen der verschiedenen Aspekte von Lesekompetenz innerhalb der DOL-Stichprobe aus vier Klassenverbänden ähnlich verteilt sind wie in den beiden Grundgesamtheiten (Drittklässler*innen und Viertklässler*innen).

² Nur bei den Schüler*innen der Kohorte 4 bis 7 zum letzten Erhebungszeitpunkt wurde der ELFE 1 – 6 nicht durchgeführt, da er nicht mehr indiziert ist.

Ko- horte	Schüler* innen-ID	Ge- schl.	Muttersp. Deutsch	Deutsch- leistg.	weit. Schule	Jgst 3	Jgst 4	Jgst 5	Jgst 6	Jgst 7
3 - 6	AIG-03	w	1	1, 0	GYM	x	x	x	x	
	AIG-09	m	1	2, 3	RS+	x	x	x	x	
	AIG-17	m	1	1, 1	GYM	x	x	x	x	
	AIG-19	m	1	1, 5	GYM	x	x	x	x	
	AIG-23	w	1	2, 0	RS+	x		x	x	
4 - 7	AIG-42	m	1	2, 0	GYM		x		x	x
	AIG-46	m	1	2, 4	RS+		x		x	x
	AIG-48	w	1	2, 4	RS+		x		x	x
	AIG-51	w	3	2, 8	RS+		x		x	x
	AIG-55	w	1	1, 6	GYM		x		x	x
	AIG-57	w	1	1, 6	RS+		x		x	x
	AIG-61	m	1	1, 9	GYM		x		x	x
	AIG-63	m	1	2, 0	RS+		x		x	x
	AIG-68	w	1	2, 9	RS+		x		x	x
	AIG-72	w	1	2, 6	GYM		x			x
	AIG-73	w	2	2, 5	GYM		x		x	x

Tab. 1: Merkmale und Kontexte der 26 untersuchten Aufnahmen. In den Zeilen sind solche Schüler*innen aufgeführt, von denen mind. eine Aufnahme bei maximaler Übereinstimmung auf Pinnell-Level 4 eingeordnet wurde (grau unterlegt). Ohne Unterlegung angekreuzt sind vorliegende Aufnahmen, die keine derart optimale Bewertung erhalten haben. Weiterhin aufgeführt sind die wichtigsten Rahmendaten zu den beteiligten Schüler*innen: Deutschkenntnisse und Deutschleistung wurden von der jew. Lehrkraft zum ersten Erhebungszeitpunkt eingeschätzt, und zwar einmal auf einer Skala von 1 (= Muttersprache) bis 4 (= kaum vorhanden) und einmal auf Schulnotenskala als Mittelwert aus acht Einzeleinschätzungen (s. o.). Bei der weiterführenden Schule steht GYM für Gymnasium und RS+ für Realschule Plus.

Von jeder/jedem der 31 Schüler*innen liegen aufgrund von vereinzelt Ausfällen innerhalb des Erhebungszeitraums im Durchschnitt drei (statt idealerweise vier) Aufnahmen vor, bei denen der DOL-Text gelesen wurde (insgesamt 99 Aufnahmen). Diese Aufnahmen wurden in Sappok, Linnemann & Stephany (2020) hinsichtlich Automatisiertheit und Akkuratheit analysiert sowie im Rahmen einer größer angelegten Ratingprozedur (insgesamt kamen 256 Aufnahmen und N = 51 Rater*innen zum Einsatz) mit Lehramtsstudierenden des Faches Deutsch einer umfassenden Bewertung unterzogen. Auf die verschiedenen Kriterien und näheren Umstände soll hier nicht erschöpfend eingegangen werden (Einzelheiten s. ebd.). Wichtig ist für den vorliegenden Kontext nur, dass jede Aufnahme von drei Rater*innen bewertet wurde. Dabei wurde u. a. die vierstufige NAEP-Skala zur globalen Erfassung von Leseflüssigkeit (Pinnell et al. 1995, Daane et al. 2005) angewendet, und zwar in der von Rosebrock & Nix (2015) für das Deutsche adaptierten Form³. Die Inter-Rater-Reliabilität erwies sich als hoch (ICC = .80).

Für die vorliegende Untersuchung wurden nun aus den 99 Aufnahmen genau diejenigen ausgewählt, die von allen dreien der jeweiligen Rater*innen auf Leseflüchtigkeitslevel 4 eingestuft

³ Einziger Unterschied: Die Stufenbeschreibungen von Rosebrock & Nix (2015) sind auf narrative Texte ausgerichtet und es wird auf das Gelesene als „die Geschichte“ referiert. Dies wurde im vorliegenden Kontext durch „der Text“ ersetzt, da ein Sachtext beurteilt werden sollte.

wurden. Dies war bei 26 Aufnahmen der Fall (Tabelle 1). Diese Aufnahmen zeichnen sich also dadurch aus, dass sie im Rahmen der verwendeten Skala mit maximaler Übereinstimmung die maximal positive Bewertung erhalten haben (post hoc ICC = 1, s. Abbildung 1).

Der so definierte Auswahlprozess bringt also ein Höchstmaß an „positiver Homogenität“ bzgl. Leseflüssigkeit auf der NAEP-Skala mit sich. Damit geht ein hohes Maß an Heterogenität im Hinblick auf andere Aspekte einher. So beinhaltet die Teilmenge Aufnahmen von sämtlichen untersuchten Zeitpunkten (Jahrgangsstufe 3 bis 7) und einzelne Schüler*innen sind mit einer bis hin zu vier Aufnahmen vertreten. Tabelle 1 fasst wichtige Merkmale der untersuchten Aufnahmen in einem Schüler*innen-zentrierten Arrangement zusammen

Auch wenn bei der „Ziehung“ dieser Stichprobe der Zufall nicht unter voll kontrollierten Bedingungen gewaltet hat, kann davon ausgegangen werden, dass damit ungefähr die langfristig lesestärkere Hälfte aus den ursprünglichen vier Grundschulklassen repräsentiert ist (unter Ausblendung von Schüler*innen mit Inklusionshintergrund). Als Ergebnis der Vorarbeiten kann diese Feststellung ins Negative gewendet und als weiterführende Hypothese folgendermaßen zusammengefasst werden:

- Hypothese 1: Aus ca. 50% aller Grundschülerinnen und Grundschüler werden bis Mitte Sekundarstufe I keine flüssigen Leser*innen.

Dieser Hypothese lässt sich entgegenhalten, dass die Stufenbeschreibung von Level 4 maximal hohe Anforderungen aufführt, indem mit expressiver Interpretation auf der Ebene des prosodischen Lesens nicht nur ein sinn- und syntaxgemäßes, sondern darüber hinaus auch ein sinn-gestaltendes Lesen eingefordert wird (s. Abbildung 1). Eigene Vorarbeiten haben allerdings mehrfach gezeigt, dass, zumindest von studentischen Rater*innen, Level 4 relativ wenig streng zugeteilt wird, wenn das Referenzfeld wie hier breit ist. So muss bei der Pinnell-Skala von einem ausgeprägten Deckeneffekt gesprochen werden (Sappok & Fay 2018, Sappok, Linnemann & Stephany 2020). Diese Feststellung kann wie folgt als weiterführende Hypothese ausformuliert werden:

- Hypothese 2 a) Das Überschreiten der Schwelle zu NAEP-Level 4 bedeutet, lediglich early prosody erworben zu haben.
- Hypothese 2 b) Innerhalb von NAEP-Level 4 (advanced prosody) ist ein breites Spektrum ausdifferenzieren.

Wichtig für das weitere Vorgehen ist die Unterscheidung, dass weniger eine Stichprobe von Schüler*innen, wie im Arrangement von Tabelle 1 aus Übersichtlichkeitsgründen nahegelegt, als eine Stichprobe von Aufnahmen behandelt wird. Aus dieser Perspektive stellt das lesende Individuum ein eher nachrangiges von mehreren Merkmalen von Aufnahmen dar. Die Aufnahmenstichprobe repräsentiert somit die Grundgesamtheit von solchen „Lautleseereignissen“ (bzw. Lernständen zu einem bestimmten Zeitpunkt) aus dem „Altersraum“ bis Mitte Sekundarstufe I, die mit maximaler Eindeutigkeit auf Pinnell-Level 4 (s. Abbildung 1) eingeordnet werden können. Diese Perspektivsetzung hat diverse Konsequenzen für das methodische Vorgehen. Ein Vorteil besteht z. B. darin, dass die Lückenhaftigkeit der Datenerhebung statistisch umso weniger ins Gewicht fällt (leere Felder in Tabelle 1), je stärker das lesende Individuum und auch dessen Jahrgangsstufe in den Hintergrund gerückt werden.

3 | Eigenschaften des Textes und der Skalen

Der als Lesestimulus eingesetzte Text stammt aus einem Schülerlexikon mit der Altersangabe „ab 9 Jahren“ (Beuschel-Menze et al. 2009, Klappentext). Es handelt sich um den Eintrag zum Begriff „Dolmetscher“:

„**Dolmetscher** sind Übersetzer von einer Sprache in die andere. Während aber ein Übersetzer Texte nur schriftlich übersetzt, muss der Dolmetscher auch in mündlichen Verhandlungen übersetzen können. Ein Simultan-Dolmetscher übersetzt sogar laufend, während weitergesprochen wird. Das seltsame Wort ist aus der osmanisch-türkischen Sprache abgeleitet. Der *tilmac* war dort ein Vermittler.“ (aus „Schülerlexikon von A - Z“, Beuschel-Menze et al. 2009: 64, Herv. i. Orig.)

Der Text weist eine hohe Informationsdichte auf. Die durchschnittliche Satzlänge beträgt 9,8 Wörter, der Anteil an langen Wörtern (Buchstabenanzahl > 6) beträgt 42,8%. Es ergibt sich ein LIX-Wert (Lesbarkeitsindex nach Björnsson, zit. nach Lenhard & Lenhard 2014) von 52,6, was auf einen erhöhten Schwierigkeitsgrad hinweist. Die Ausführungen und Unterscheidungen des Textes sind vielschichtig, indem sie Merkmale einer anspruchsvollen Definition aufweisen: Im ersten Satz wird der Übersetzer-Begriff als *genus proximum* vorgestellt; im zweiten Satz wird der Gegensatz Schriftlichkeit-Mündlichkeit als *differentia specifica* zur Abgrenzung von einem damit implizit revidierten Übersetzerbegriff eingeführt, was als ziemlich problematisch gelten kann. Im dritten Satz wird über das seltene Wort „simultan“ noch einmal eine Ebene tiefer ausdifferenziert. Im vierten und fünften Satz wird eine etymologische Einordnung vorgenommen, wobei die Lautung des Ursprungsbegriffs „tilmac“ nicht eindeutig geklärt wird. Vor diesem Hintergrund muss der Text als *kein* guter Beitrag zu einem Schülerlexikon „ab 9 Jahren“ (Beuschel-Menze et al. 2009, Klappentext) angesehen werden. Gerade vor diesem Hintergrund aber, nämlich weil er so auch für Siebtklässler*innen noch schwierig genug ist, kann der Text als ein gut geeigneter Stimulus für die Diagnostik von advanced prosody gelten.

Dazu lässt sich festhalten, dass der Text einige bemerkenswerte prosodische Herausforderungen enthält. Die Verhältnisse auf Ebene der Informationsstruktur sind bei aller Komplexität relativ eindeutig gelagert, sodass sich ein Mitvollziehen dieser Verhältnisse seitens des Lesenden auf der Ebene der fokussierenden stimmlichen Hervorhebung bestimmter Wörter niederschlagen kann (z. B. „Dolmetscher sind Übersetzer von EINER Sprache in die andere.“ vs. „Dolmetscher sind Übersetzer von einer SPRACHE in die andere.“). Dies ist als besonders anspruchsvoller Aspekt von prosodischem Lesen bzw. als potenzieller Indikator für komplexe Aspekte von Textverstehen anzusehen: „By highlighting key information, good prosody in the reading of linguistic focus features can signal the ongoing construction of a discourse structure during fluent reading“ (Schwanenflugel, Westmoreland & Benjamin 2013, weitere Ausführungen hierzu in Sappok, Linnemann & Stephany 2020).

Die untersuchten Audioaufnahmen sind mit einem Headsetmikrofon vom Typ Sennheiser ME 3-ew entstanden, das mit einem digitalen Aufnahmegerät vom Typ Olympus LS-11 verbunden war. Das Format der Audiodateien ist überall Mono, 44.1-kHz, WAV 16-bit. Die Dauern der Aufnahmen liegen zwischen 20 und 34 Sekunden. Hierfür sind Unterschiede im Eigentempo der Schüler*innen sowie Stockungen und Wiederholungen (zwischen 0 und 6 Lesefehler pro Aufnahme) verantwortlich zu machen. Für die Audioanalyse im Rahmen der nun vorgestellten Ratingprozedur wurden sämtliche Aufnahmen an den Satzgrenzen in Sequenzen geschnitten, die im Folgenden als Audiophrasen bezeichnet werden sollen. Der längste Satz (Satz 2) wurde an

der Kommasetzung auch noch einmal geteilt, um die Ähnlichkeit in der Länge der somit resultierenden 6 Audiophrasen je Aufnahme zu erhöhen (Länge zwischen 17 und 22 Silben für die ersten fünf Phrasen; einzig die sechste Phrase (der letzte Satz) weicht mit einer Länge von 9 Silben nennenswert nach unten ab).

Alle Rater*innen hatten damit dieselben $26 * 6 = 156$ Audiophrasen zu beurteilen, allerdings nach unterschiedlichen Kriterien und in individuell randomisierter Abfolge. Es kamen drei Skalen zum Einsatz, mit denen drei gleichgroße Gruppen à 13 Rater*innen gearbeitet haben. Sämtliche Ratingsitzungen fanden gleichzeitig in einem Zeitfenster von ca. 60 min statt, und zwar an Rechnerarbeitsplätzen in einer mittels der Softwareumgebung Praat/ExperimentMfC (Boersma & Weenink 2020) automatisierten Ablauforganisation auf einer graphischen Benutzeroberfläche. Die Rater*innen erhielten ein Daten- und Instruktionpaket zum Download und arbeiteten zuhause. Sie waren angewiesen, möglichst hochwertige Kopfhörer einzusetzen. Über Modalität und Qualität der von den Rater*innen verwendeten Audiotechnik konnte darüber hinaus keine Kontrolle ausgeübt werden.

Der mausgesteuerte Bildschirm- und Audio-Ablauf bestand darin, dass alle Rater*innen nach einer teamspezifischen Start-Instruktionfolie (s. Abbildung 4) zuerst den ersten Satz des DOL-Textes in 26 Audiophrasen in individuell randomisierter Reihenfolge präsentiert bekamen, dann den ersten Teil des zweiten Satzes in 26 Audiophrasen, dann den zweiten Teil des zweiten Satzes usw (s. Tabelle 2).

Der Start einer einzelnen Audiophrasenpräsentation bestand in einem automatischen Abspielen der Phrase. Dazu wurde die per Mausklick zu wählende Option gegeben, die Phrase noch einmal zu hören (bis zu drei Wiederholungen waren möglich). Daneben war die entsprechende Fragestellung aufgeführt, zusammen mit Buttons zur Beantwortung. Wenn eine Beantwortung erfolgt war, konnte sie korrigiert oder aber es konnte per Mausklick zur nächsten Audiophrasen-Präsentation übergegangen werden, die dann wieder mit dem automatischen Abspielen startete. Die Rater*innen hatten in mehreren Online-Seminarsitzungen eine basale Einführung in den Umgang mit der Software Praat und den Kontext „Hörexperiment“ bekommen, wobei zum Kennenlernen des Textes auch schon andere DOL-Aufnahmen aus dem LAUDIO-Korpus zum Einsatz gekommen waren.

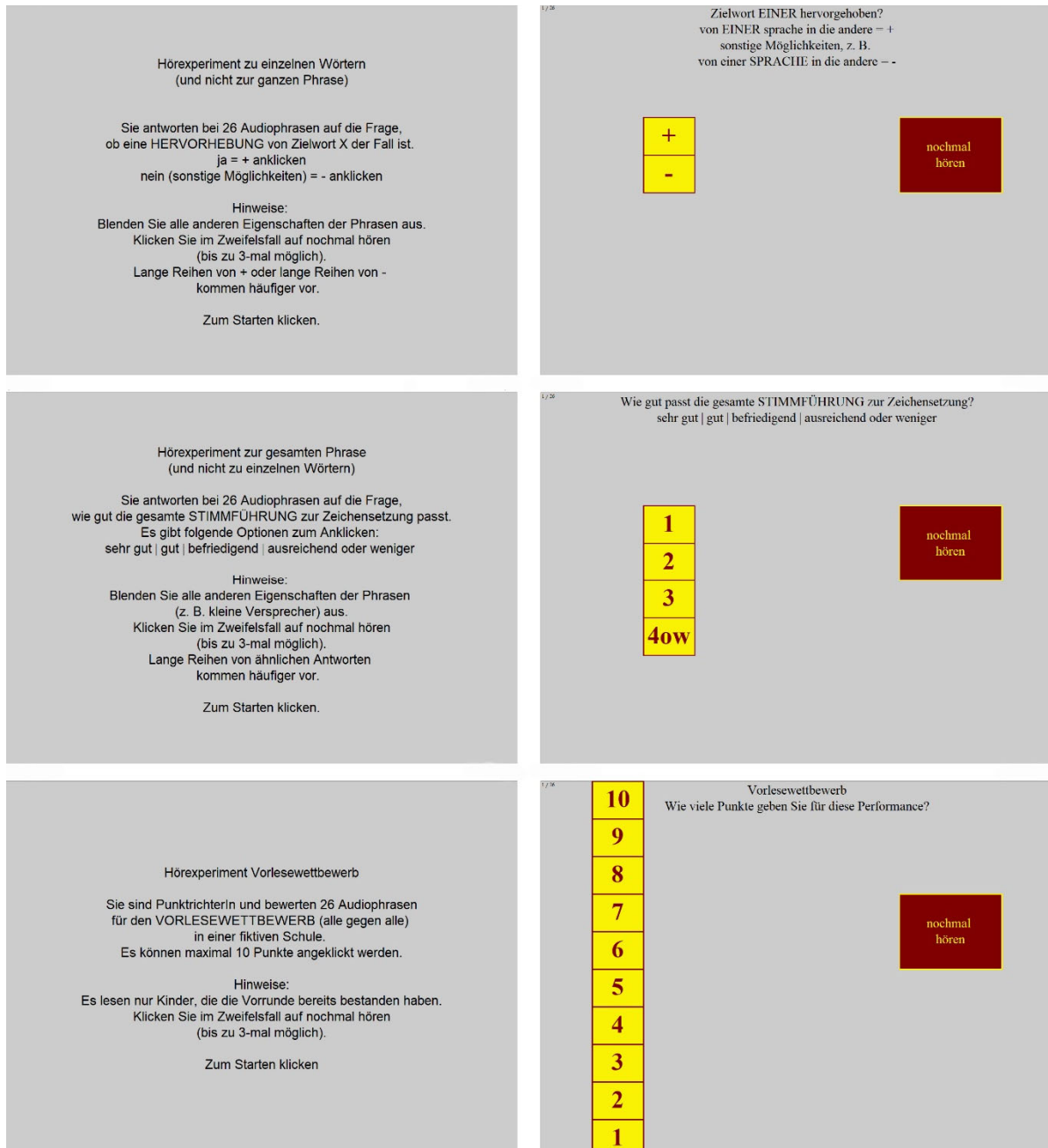


Abb. 4: Die unter ansonsten identischen Bedingungen zum Einsatz gekommenen Ratingskalen in Form der Bildschirmanzeige, mit der die jew. Sitzungen starteten und der jew. ersten Bildschirmanzeige zur ersten Audiophrasen-Präsentation.

Die Instruktionen A und B sind mit dem Ziel formuliert, bei den Rater*innen die Konzentration auf den erfragten Aspekt zu maximieren. Die Instruktion C ist im Gegensatz dazu mit dem Ziel formuliert worden, ein maximal holistisches Urteil zu erhalten. Die Fokussierung auf HERVORHEBUNG in Skala A zielt auf das Konzept der phonetischen Prominenz auf Wortebene ab. Bewusst verzichtet wurde hier auf die zusätzliche Fokussierung auf die semantische bzw. informationsstrukturelle Motiviertheit der anvisierten Prominenzverhältnisse, weil dies zu umständlich zu vermitteln und außerdem zu befürchten ist, dass ansonsten der Faktor des Rater*innen-

eigenen Textverständnis störenden Einfluss gewinnen würde (im Sinne einer Rater*in-eigenen „Informationsstruktur-Kompetenz“). Im Ablauf wurde dann je Phrase die Hervorhebung eines bestimmten Wortes im Gegensatz zu dessen Nicht-Hervorhebung erfragt. Zusätzlich wurde jew. eine Gegenbeispielvariante angegeben. Die zugrundeliegende Bestimmung der Zielwörter war davon geprägt, dass deren Hervorhebung im Zentrum eines Prominenzmusters steht, das maximal im Sinne der Informationsstruktur ist (Muster „+ Informationsstruktur“). Im Einzelnen handelt es sich um folgende Muster und damit Angaben in den betreffenden Positionen der aktuellen Anzeige auf der Benutzeroberfläche (vgl. Abb. 4 A rechts):

Phrase	Muster „+ Informationsstruktur“	Muster „- Informationsstruktur“
1	von EINER sprache in die andere	von einer SPRACHE in die andere
2-1	nur SCHRIFTLICH übersetzt	nur schriftlich ÜBERSETZT
2-2	auch in MÜNDLICHEN verhandlungen	auch in mündlichen VERHANDLUNGEN
3	während WEITER gesprochen wird	während weiter GESPROCHEN wird
4	das seltsame WORT ist	das seltsame wort ist
5	der TILMAC war	der tilmac war

Tab. 2: „Sinngestaltende Prosodie“. Die auf der binären Skala A erfragten Angaben zu den in der jeweiligen Audiophrase aktualisierten Prominenzverhältnissen.

Bei den Phrasen 4 und 5 wurde wegen mangelnder klarer Alternativen auf Ebene der Informationsstruktur für die Gegenüberstellung eine hervorhebungsneutrale prosodische Gestaltung und damit das anvisiert, was alltagssprachlich „monotones“ Lesen genannt wird. Die Buchstabenfolge <tilmac> wird als eine Art Pseudowort angesehen, von dem auch ohne eindeutige Lautung klar sein sollte, welche Rolle es im Satz spielt. Zusammenfassend lässt sich festhalten, dass es sich bei Skala A um eine exkludierende („sonstige Möglichkeiten“) binäre Skala handelt, die zunächst als zweistufige Nominalskala (Reliabilitätsprüfung über Light’s Kappa) behandelt wird (s. Tabelle 3).

Die Fokussierung STIMMFÜHRUNG bei Skala B bezieht sich auf die Gesamtkontur der Intonationsphrase; mit der zusätzlichen Fokussierung „satzzeichengemäß“ werden die von syntaktischen Zusammenhängen bestimmten Intonationsverhältnisse, v. a. Deklination (Pompino-Marschall 2009) fokussiert. Da die Satzzeichenverhältnisse bei allen Phrasen recht klar sind, wurde erwartet, dass eine Rater*innen-eigene „Satzzeichenkompetenz“ hier weniger störend sein würde als eine Rater*innen-eigene „Informationsstruktur-Kompetenz“ bei Skala A. Um einer durch den Begriff „Satzzeichen“ nahegelegten Beschränkung der Fokussierung auf das Ende oder verstärkt die zweite Hälfte der Audiophrase entgegenzuwirken, hieß die Instruktion, die bei jeder Präsentation erschien: „Wie gut passt die *gesamte* STIMMFÜHRUNG zur Zeichensetzung?“ (vgl. Abbildung 4 B rechts). Zusammenfassend lässt sich festhalten, dass es sich bei Skala B um eine trunkierte („ausreichend oder weniger“) Likert-Skala auf Ordinalniveau handelt, die im Folgenden aus methodischen Gründen als vierstufige Intervallskala behandelt wird (vgl. hierzu die Diskussion in Rietveld & Chen (2006: 303f.)).

Die Fokussierung VORLESEWETTBEWERB bei Skala C bezieht sich auf den oben formulierten Heurismus 1. Vorausgesetzt wird, dass schon eine isolierte Audiosequenz von wenigen Sekunden Dauer ausreichend diagnostische Information enthält, wenn es um die Kodierung einer globalen Vorlesekompetenz geht. Damit ist auch mit einem erhöhten Einfluss der Rater*innen-eigenen Zuhörkompetenz zu rechnen. Über die Hinweise „(alle gegen alle) in einer fiktiven Schule“ (vgl. Abb. 4 C links) sollte ein Casting-Szenario wie aus den Medien bekannt suggeriert werden. Dabei sollte dem Einfluss entgegengewirkt werden, dass die teilweise gut hörbaren Altersunterschiede bei den Audiophrasen in die Beurteilung als Gewichtung einfließen. Die Benutzeroberfläche bei den einzelnen Präsentationen war im Gegensatz zu A immer dieselbe und lautete „Wie viele Punkte vergeben Sie für diese Performance?“ (vgl. Abbildung 4 C rechts).

Drei der 39 Rater*innen lieferten technisch fehlerhafte Daten. Bei der Analyse von Inter-Rater-Reliabilität wurden dann zwei weitere Rater*innen auffällig, da sie in einer ihrer Skalen-Phrasenkombination immer denselben Wert angeklickt hatten, was zu einer Fehlermeldung wegen Nullvarianz führte. Die betroffenen Datensätze wurden manuell entfernt. Weitere Maßnahmen zu Identifikation und Ausschluss unkooperativer oder überforderter Rater*innen fanden nicht statt.

	A	A	B	C	A	B	C
Phrase	Light's	ICC _{tcs}	ICC _{tcs}	ICC _{tcs}	ICC _{tca}	ICC _{tca}	ICC _{tca}
1	.28	.26	.33	.59	.82	.86	.95
2-1	.29	.31	.39	.55	.86	.89	.93
2-2	.27	.28	.25	.47	.84	.81	.92
3	.21	.23	.30	.57	.79	.85	.94
4	.11	.10	.36	.70	.59	.88	.96
5	.16	.19	.33	.60	.75	.87	.94
<i>Median</i>	.24	.25	.33	.58	.81	.87	.94

Tab. 3: Inter-Rater-Reliabilität der drei Skalen im Spiegel unterschiedlicher Indizes. Werte bei ICC_{tcs} < .5 lassen relativ eindeutig auf Verbesserungsbedarf bei der Skalenkonstruktion schließen.

Aufgrund der grundsätzlich symmetrischen und stark kontrollierten Bedingungen der Gesamt-Ratingprozedur konnten besonders differenzierte Analysen der Inter-Rater-Reliabilität durchgeführt werden. Für die binäre Skala A wurde Light's Kappa (nach Hallgren 2012) eingesetzt, ein Parameter für nominal skalierte Faktoren, dessen Algorithmus für jede mögliche Rater*innen-Paarung innerhalb eines Phrasensets einen einfachen Kappa-Wert und aus all diesen dann den Mittelwert liefert. Analysen wurden in der Softwareumgebung R (R Core Team 2014) mithilfe des Paketes „irr“ (Gamer, Lemon & Singh 2019) durchgeführt. Dabei liegen, wie bei den anderen im folgenden verwendeten Indizes, die Ergebnisse i. d. R. zwischen 0 (keine Übereinstimmung) bis 1 (maximale Übereinstimmung).

Die Resultate für Skala A müssen teilweise als nicht ausreichend für eine vollwertige weiterführende Analyse dieser Daten angesehen werden, zumal hier der naheliegendste nächste Schritt in einer Reduktion auf eine binäre Ausprägung pro Audiophrase besteht (Item gelöst vs. nicht gelöst). Die Phrasen 1 bis 3 erreichen Light's Kappa Werte zwischen .2 und .3 und damit das

Level „fair agreement“ (nach Hallgren 2012). Die Phrasen 4 und 5 liegen noch einmal merklich niedriger. Dies könnte darauf zurückzuführen sein, dass bei 4 und 5 im Ablauf keine eindeutige Hervorhebungsalternative aufgezeigt war (siehe Abbildung 4).

Die Inter-Rater-Reliabilität der Daten aller Skalen wurde außerdem mit ICC-Analysen geprüft. Es zeigte sich für Skala A eine starke Übereinstimmung mit Light's Kappa, sodass der unmittelbare Vergleich zu B und C möglich ist. Hierbei spielen zwei Perspektiven eine Rolle. Um Aufschlüsse zur Skalenkonstruktion und zu den Hörkompetenzen der Rater*innen zu erhalten, wurden die Parameter „twoway, consistency, single unit“ (= ICC_{tcs}) gesetzt, was zu konservativen Ergebnissen im Hinblick auf die hypothetische Reliabilität der Daten führt, aber feiner differenziert. Bei B zeigt sich bis auf Phrase 2-2 durchgängig eine etwas höhere Tendenz als bei A. Bei Skala C hingegen zeigt sich klar ein insgesamt höheres Niveau. Diese Verhältnisse lassen sich wie folgt zusammenfassen: $IRR.A < IRR.B \ll IRR.C$.

Mit Perspektive auf die weiterführende Analyse der Daten bzgl. der Lesekompetenzen der Schüler*innen ist zu berücksichtigen, dass das Setting mit sehr vielen Rater*innen darauf ausgelegt ist, die rohen Ratingdaten zu Audiophrasen-spezifischen Mittelwerten zusammenzufassen (s. nächstes Unterkapitel). Ausschlaggebend ist deshalb nach Hallgren (2012; vgl. Sappok & Fay 2018) die Analyse, bei der die Parameter „twoway, consistency, average“ (ICC_{tca} in Tabelle 3) gesetzt sind, was zu progressiven Ergebnissen im Hinblick auf die hypothetische Reliabilität der Daten führt. Alle ermittelten ICC_{tca} -Werte für die Skalen B und C sind größer als .75 und genügen damit hohen Ansprüchen („excellent“ nach Hallgren 2012).

Festgehalten werden kann somit, dass die Skala C anhand kurzer Audiosequenzen maximal reliable Daten für die weiterführende Analyse liefert. Die Inter-Rater-Reliabilität dürfte auch einer merklichen Verringerung der Rater*innen-Anzahl unter ansonsten günstigeren Bedingungen (z. B. Präsenz, Schulung, hochwertige Audiotechnik) standhalten. Geringer ist die Inter-Rater-Reliabilität für die Skala B einzustufen, was in der vorliegenden Untersuchung durch eine hohe Rateranzahl tendenziell aufgefangen wird. Die Skala A muss als wenig reliabel eingestuft werden, wobei Phrase 4 und Phrase 5 als besonders wenig reliabel identifiziert wurden. Die Ratingdaten A für diese Phrasen wurden für den zweiten Schritt der weiterführenden Analyse (Zusammenfassung zu Aufnahmenkennwerten, s. u.) ausgeschlossen.

4 | Ergebnisse

Für die Analyse werden die Ratingdaten in zwei Schritten durch Mittelwertbildung kondensiert und weiteren Merkmalen gegenübergestellt. Der erste Schritt fokussiert die 156 Audiophrasen (Mittelwert aus max. 13 Ratings pro Skala und Audiophrase), der zweite Schritt fokussiert die 26 Gesamtaufnahmen (Mittelwert aus max. 78 Ratings pro Skala). Im Zentrum stehen dabei die Ratings der Skala C und die Frage nach Hinweisen dazu, worauf sich eine mehr oder weniger hohe C-Punktzahl einer Aufnahme im Einzelnen zurückführen lässt.

4.1 | Ergebnisse auf Ebene der Audiophrasen

Pro Audiophrase liegen bzgl. der C-Skala 13 Einzelratings vor, die zu einem Mittelwert zusammengefasst wurden. Es ergibt sich eine Verteilung von 156 Werten mit global (s. Tabelle 4, unterste Zeile): $M = 6,74$, $SD = 1,48$, Range = 3,67 bis 9,42. Abbildung 5 zeigt diese Werte im Ranking. So erweist sich, dass die Skala breit und relativ gleichmäßig ausgenutzt wurde.

Phrase	Wörter	Silben	Einzel- ratings	C	B	A	E	S
Num- mer	Anzahl	Anzahl	Anzahl	M (SD)	M (SD)	Σ	M (SD)	M (SD)
1	9	18	13	7,10 (1,54)	7,64 (1,25)	3	- 0,04 (0,20)	4,57 (0,69)
2-1	8	17	13	7,08 (1,41)	7,46 (1,37)	16	- 0,46 (0,65)	4,40 (0,70)
2-2	9	20	13	7,07 (1,21)	7,68 (1,08)	18	- 0,38 (0,64)	4,59 (0,80)
3	9	22	13	5,53 (1,35)	7,11 (1,20)	16	- 0,77 (0,65)	3,86 (0,83)
4	10	20	13	6,76 (1,42)	7,80 (1,20)	-	- 0,62 (0,80)	3,96 (0,65)
5	6	9	13	6,90 (1,43)	7,22 (1,07)	-	- 0,12 (0,33)	3,46 (0,73)
Gesamt	51	106	78	6,74 (1,48)	7,65 (1,23)	53	- 0,40 (0,63)	4,14 (0,83)

Anmerkung: Die A-Daten zu den Phrasen 4 und 5 wurden wegen mangelnder Reliabilität ausgeschlossen.

Tab. 4: Kennwerte der Audiophrasen (N = 156) auf Phrasenebene und insgesamt.

Vergleichbar wurde mit den Ratings der Skala B verfahren. Erst wurden die als Schulnoten erfragten Daten in Punktzahlen umgerechnet („sehr gut“ = 4 Punkte bis „ausreichend oder weniger“ = 1 Punkt; $M = 3,06$ (i. e. „gut“), $SD = 0,49$, Range = 1,62 bis 3,92). Multipliziert man die Punktzahlen mit dem Faktor 2,5, ergeben sich die besser auf C beziehbaren B-Kennwerte auf einer Skala von 1 bis 10 (B global: $M = 7,65$ und $SD = 1,23$ und Range = 4,04 bis 9,81 (siehe Tabelle 4, unterste Zeile)). Für die rechnerischen Analysen ist diese Transformation irrelevant. Der höhere Mittelwert bei B im Gegensatz zu C bei geringerer Streuung zeigt, dass die B-Skala weniger gleichmäßig, sondern eher im höheren Bereich ausgenutzt wurde. Die nach rechts spitz zulaufende Punktwolke in Abbildung 5 zeigt eine Tendenz abnehmender Abweichung zwischen B und C, d. h. je höher C, desto enger geht C mit einer „insgesamt satzzeichengemäßen Stimmführung“ einher. Hier spielt augenscheinlich auch ein Deckeneffekt eine Rolle.

Auch die Ratings der Skala A wurden zu Mittelwerten zusammengefasst. Die Fragestellung nach der Hervorhebung eines Wortes in einer Phrase im Gegensatz zu allen anderen möglichen Hervorhebungsverhältnissen lässt allerdings keine sinnvolle Interpretation der Anteilhaftigkeit zu,

die mit diesen Mittelwerten suggeriert wird. Deshalb wurde gerundet, d. h. war der Mittelwert aus den 13 Ratings $\geq 0,5$, erhielt die jew. Audiophrase die Ausprägung $A = 1$ (= Hervorhebung informationsstrukturgemäß), sonst 0. Da die Werte zu Phrase 4 und 5 aus der Analyse ausgeschlossen wurden, liegen 104 Ausprägungen vor, global hiervon 53-mal 1 (s. Tabelle 4, unterste Zeile). So erscheinen in Abbildung 5 zwei Kettenstrukturen auf der 0- und auf der 1-Linie, die allerdings von fehlenden Werten (Phrase 4 und Phrase 5) durchsetzt sind. Ein Zusammenhang mit den C-Daten zeichnet sich dahingehend ab, dass etwa ab $C > 7$ eine höhere „Kettendichte“ auf der 1-Linie als auf der 0-Linie ersichtlich ist, d. h. mit hoher C-Punktzahl geht meist auch eine informationsstrukturgemäße Worthervorhebung einher.

Als weitere Merkmale werden auf der Ebene der Audiophrasen die Fehleranzahl E („Errors“) und die Sprechgeschwindigkeit S [Silben/sek] einbezogen (Korrelate von Akkuratheit und fortgeschrittener Automatisierung). Als Fehler gezählt wurden Auslassungen, Ersetzungen und Hinzufügungen auf Wortebene (Daane et al. 2005). Da es sich bei den Audiophrasen um kurze Sequenzen handelt, sind die meisten fehlerfrei (106-mal 0 Fehler). Von den verbleibenden 50 Aufnahmen weisen 38 einen und 12 zwei Fehler auf. Für Abbildung 5 (rote Kettenstrukturen) wurden diese Werte durch $(E * (-1))$ ersetzt, um die positive Konnotation der Y-Achse aufrecht zu erhalten, und 0 Fehlerpunkte wurden herausgelöscht, damit sie nicht auch als rote Punkte und damit optisch als Minuspunkte in Erscheinung treten. Die unausgewogene Verteilung schlägt sich in der auffälligen Mittelwert-Streuungskombination von $M = 0,4$ und $SD = 0,63$ nieder (s. Tabelle 4, unterste Zeile). Abbildung 5 lässt klar erkennen, dass die Fehler schlüssigerweise fast ausschließlich im unteren Bereich von C auftreten.

Die Sprechgeschwindigkeit S wurde in Sprechsilben/Sekunde gemessen. Für die fehler- und pausenfreien Audiophrasen wurde die absolute Dauer von Sprechbeginn bis Sprechende gemessen und mit der Silbenanzahl der Textphrase verrechnet. Jede von Fehlern betroffene Phrase wurde auf zusätzliche oder fehlende Sprechsilben sowie merkliche Pausen geprüft und das Ergebnis mit der Silbenanzahl der jew. Textphrase und dann nach dem beschriebenen Muster mit der absoluten Dauer, ggf. abzüglich Pausen, verrechnet [tats. Sprechsilben/sek]. Es ergaben sich die globalen Kennwerte von $M = 4,14$ und $SD = 0,83$ und $Range = 2,02$ bis $6,73$ (s. Tabelle 4, unterste Zeile). Gestört wird S durch Stockungen im Redefluss, wie sie oft mit Fehlern einhergehen und nicht als Pausen in Erscheinung treten, und zwar im Hinblick auf das auf Konstruktebene avisierte latente Personenmerkmal Eigentempo.

Zum Begriff Eigentempo ist ein kurzer Exkurs angebracht: Ein intraindividuell konstantes Eigentempo kann als prosodisches Gestaltungsmittel bei fortgeschrittenem Vorlesen gelten und ist auch phonetisch klar bestimmbar, allerdings nur bei fehler- und stockungsfreien Passagen. Eine transparentere Maßeinheit wäre hier $(S * 60) / 2$, also Doppelsilben pro Minute. Unter den verallgemeinernden Annahmen, dass im Durchschnitt jede zweite Silbe betont ist und dass ein Wort aus durchschnittlich zwei Silben besteht, lässt sich $(S * 60) / 2$ ungefähr auf die Einheit beats per minute (BPM) bzw. words per minute (WPM) beziehen. So ergibt sich für BPM oder WPM global: $M = 124$, $SD = 24$, $Range = 61$ bis 202 . Über die entsprechenden Schwellwerte zu den Tempo-Begriffen aus der Musik ist im Hinblick auf ein latentes Personenmerkmal Eigentempo festzuhalten, dass die Messungen zu den 156 Audiophrasen den Intervallen andante (76-108 BPM), moderato (108-120 BPM), allegro (120-168 BPM) und presto (168-200 BPM) zugeordnet werden können. Die Transformation ist rechnerisch irrelevant und wird in den Darstellungen nicht berücksichtigt, weil Silben/sek auf einer 10er-Skala graphisch gut eingeordnet werden

können. Abbildung 5 zeigt einen Zusammenhang zwischen S und C dahingehend, dass zwischen den auf Skala C ganz schwachen und ganz starken Phrasen ein Anstieg von ca. S = 4 auf S = 5 einhergeht.

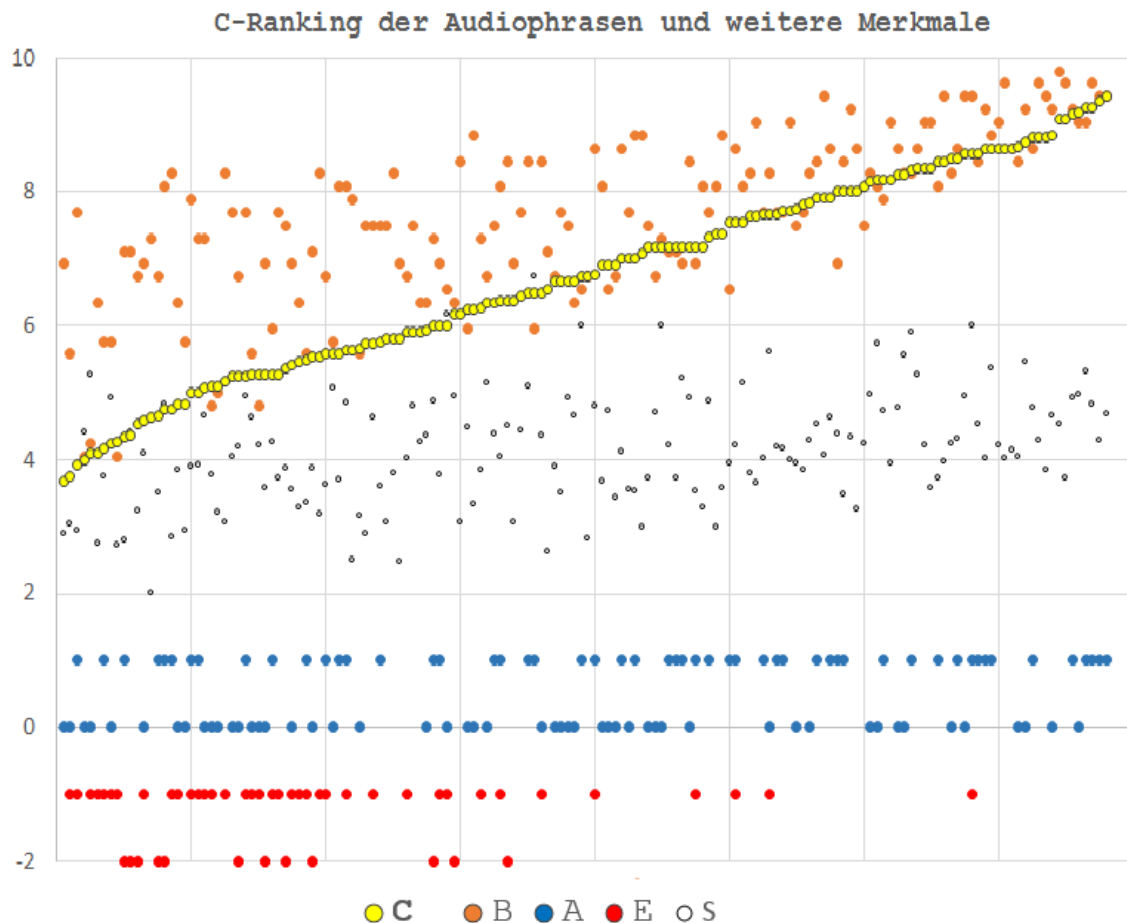
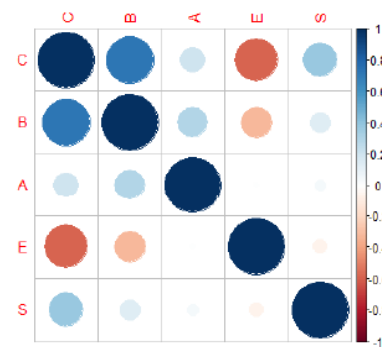


Abb. 5: Datenübersicht auf Ebene der 156 Audiophrasen im Ranking nach C. Erfasste Merkmale zu n = 156 Audiophrasen: C = Rating-Mittelwerte („Vorlesewettbewerb“), B = Rating-Mittelwerte („Stimmführung“), A = Rating-Mittelwerte („Hervorhebung“), E = - Fehlersumme („Errors“), S = Silben/sek („Speed“).

Die nach Textphrasen aufgeschlüsselten Analysen (s. Tabelle 2) lassen Aussagen zur oben angesprochenen Rater*innen- bzw. Konstruktions-Perspektive zu. Auf Textphrasenebene sind spaltenweise Abweichungen unerwünscht, weil sie auf Einflüsse schließen lassen, die Inhalt oder Stellung der Phrasen auf die Bewertung durch die Rater*innen haben. Der Faktor Phrasenlänge ist bis auf Phrase 5 wohlkontrolliert und Skala C und B weisen kaum Abweichungen auf. Bei A fällt nur Phrase 1 auf, bei der den Ratings zufolge in nur 3 von 26 Fällen informationsstrukturgemäß hervorgehoben wurde. Allerdings zeigen sich die Phrasen stark unterschiedlich anfällig für Fehler, was auf unterschiedliche Leseschwierigkeit hinweist und nicht überraschend ist. Hinsichtlich der Sprechgeschwindigkeit ist einzig Phrase 5 auffällig. Somit spricht genügend Evidenz dafür, die in Tabelle 4 vorgestellten Daten für die nachfolgenden Analysen ohne weiteren Datenausschluss zusammenzufassen.

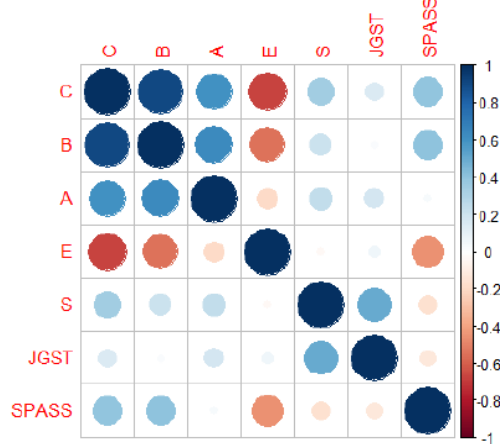
a) Phrasen

	C	B	A	E	S
C	1				
B	.73	1			
A	.21	.30	1		
E	-.58	-.32	.01	1	
S	.37	.14	.04	-.07	1



b) Aufnahmen

	C	B	A	E	S	JGST	SPASS
C	1						
B	.90	1					
A	.60	.63	1				
E	-.67	-.54	-.19	1			
S	.34	.21	.24	-.03	1		
JGST	.15	.02	.18	.07	.50	1	
SPASS	.40	.40	.03	-.45	-.16	-.12	1



Anmerkungen: Rating-Mittelwerte C = „Vorlesewettbewerb“, B = „Stimmführung“, A = „Hervorhebung“, E = Fehlersumme („Errors“), S = Silben/sek („Speed“), Jahrgangsstufe (JGST) und SPASS („Engagement“, s. Sappok, Linnemann & Stephany 2020).

Abb. 6: Korrelationskoeffizienten und Visualisierung:

- a) für n = 156 Audiophrasen
- b) für n = 26 Aufnahmen

Die in Abbildung 5 visualisierten Verhältnisse sollen nun über Korrelationen genauer in den Blick genommen werden. Hierfür wurden die fehlenden A-Werte für Phrase 4 und 5 durch den Aufnahme-internen Mittelwert aus den vorliegenden Phrasen 1 bis 3 ersetzt, um sie rechnerisch zu neutralisieren. Abbildung 6 a) zeigt die Korrelationsmatrix für die auf Ebene der Audiophrasen einbezogenen Faktoren und die Visualisierung der Koeffizienten. Die Ergebnisse fügen sich zu einem klaren Bild. Hinsichtlich der Frage nach der Einflussgröße auf C zeigt sich eine Rangfolge von $B > E > S > A$, wobei alle untersuchten Faktoren, auch der niedrigste von $r = .21$ für A, auf $p < .01$ -Niveau signifikant sind (p-Werte sind deshalb in Abb. 6 nicht mit dargestellt).

Damit ist die Analyse auf der Ebene der Audiophrasen abgeschlossen. Sie mündet vor allem in die Frage nach der Trennschärfe zwischen C und B (Korrelation zu hoch) und die Frage nach der Rolle von Faktor A (Korrelation zu niedrig). Ein differenzierteres Bild zeigt sich nach der Zusammenfassung der Phrasenmittelwerte zu Aufnahme-Mittelwerten (bis auf S, das auf Ebene der

Aufnahmen neu berechnet wurde, da so der Faktor unterschiedliche Phrasenlängen neutralisiert werden konnte).

4.2 | Ergebnisse auf Ebene der Aufnahmen

Abbildung 6 b) zeigt die Korrelationsmatrix für die auf Ebene der Gesamtaufnahmen einbezogenen Faktoren und die Visualisierung der Koeffizienten. Im Vergleich mit der Phrasenebene zeigen sich insgesamt klar höhere r -Werte. Grund ist, dass die Zusammenführung der 6 Audiophrasen zu einer Aufnahme Störvariablen zunehmend unter Kontrolle bringt.

Auf der Ebene der Aufnahmen können nun auch Merkmale einbezogen werden, die auf Phrasenebene nicht berücksichtigt wurden, weil sie bei allen Phrasen einer Aufnahme gleich ausgeprägt sind (z. B. Engagement („SPASS“ in Abbildung 6, nähere Erläuterung unten; diese Daten wurden bereits in der Vorgängeruntersuchung (Sappok, Linnemann & Stephany 2020) erhoben).

Zur Schulform kann ausgesagt werden, dass acht der 16 beteiligten Kinder auf ein Gymnasium (GYM) und acht in eine Realschule Plus (RS+) übergegangen sind (s. Tabelle 1), die Ausprägungen sind also gleich verteilt. Weitere Aufschlüsse werden im vorliegenden Kontext von diesem Faktor aber nicht erwartet, zumal auch Aufnahmen aus der Grundschule einbezogen sind. Zum Geschlecht kann festgestellt werden, dass 11 Aufnahmen von Schülern und 15 von Schülerinnen stammen. Der C -Mittelwert für die Schüler ist $M = 6,96$, $SD = 0,72$, Range = 5,60 bis 8,12, für die Schülerinnen ergibt sich: $M = 6,59$, $SD = 1,14$, Range = 4,97 bis 8,43. Der Leistungsunterschied ist nicht erheblich, die Streuung ist bei den Schülerinnen bedeutend höher, was auf unkontrollierte Einflüsse beim Zustandekommen der Aufnahmenstichprobe zurückgeführt wird. Eine sinnvolle Interpretation kann nicht gefunden werden. Das Geschlecht ist im vorliegenden Kontext als wenig C -relevant anzusehen und wird im Folgenden nicht berücksichtigt.

Im Hinblick auf Alter bzw. Jahrgangsstufe der lesenden Schüler*innen zum Zeitpunkt der Aufnahme (vgl. Tabelle 1) wurde Jahrgangsstufe (JGST) der Variable „Alter in Tagen zum Aufnahmezeitpunkt“ (ebenfalls erhoben) vorgezogen. JGST und Alter korrelieren mit $r = .97$, doch JGST ist an sich relevanter für didaktische Interpretationskontexte. Zu den entsprechenden ganzzahligen Werten (vgl. Tabelle 1) wurde 0,5 hinzuaddiert, da die Aufnahmen stets zum Halbjahreswechsel des betreffenden Schuljahrs gemacht wurden (Transformation für die Analyse irrelevant). Bei JGST ist $M = 6,38$, $SD = 1,14$, Range = 3,5 bis 7,5.

Bei „SPASS“ handelt es sich um diejenige Skala, die in der Vorgängeruntersuchung (Sappok, Linnemann & Stephany 2020) nach „NAEP-Leseeflüssigkeit“ die zweitreliabelste war. In der seinerzeit eingesetzten Ratingprozedur waren die Aufnahmen analog zu NAEP-Leseeflüssigkeit (hier immer 4, vgl. Abbildung 1) nach dem Kriterium zu beurteilen gewesen: „Spaß: Wie gut hat dem Kind die ‚Vorlesearbeit‘ gefallen?“, und zwar auf einer Skala von „gar nicht“ = 1 bis „sehr gut“ = 4. In der Längsschnittperspektive hatte sich gezeigt, dass SPASS eher ein über die Zeit intraindividuell konstantes Personenmerkmal ist, als dass in den untersuchten Biographien die SPASS-Ausprägung mit der individuellen Entwicklung bei NAEP-Leseeflüssigkeit ansteigt (ebd.). Bei SPASS ist $M = 3,02$, $SD = 0,56$, Range = 2 bis 4.

Im Hinblick auf die Korrelationsverhältnisse zeigt der Vergleich von Abbildung 6 a) und 6 b) eine etwas andere Rangfolge bei den einzelnen Einflussfaktoren auf C: Abb. 6 a) $B > E > S > A$; Abb. 6 b) $B > E > A > S$.

Außerdem muss die Höhe der Korrelation $C \sim B$ in 6 b) als Anzeichen für mangelnde Trennschärfe in den Blick genommen werden.

S korreliert in beiden Perspektiven auffällig schwach mit C. Weiterhin korreliert S, ähnlich gering, mit JGST (siehe Abbildung 6 b). SPASS korreliert mit C in merklichem, aber ähnlich geringem Maße wie S. Vergleichbar korreliert SPASS mit B und E, aber nicht mit A und S.

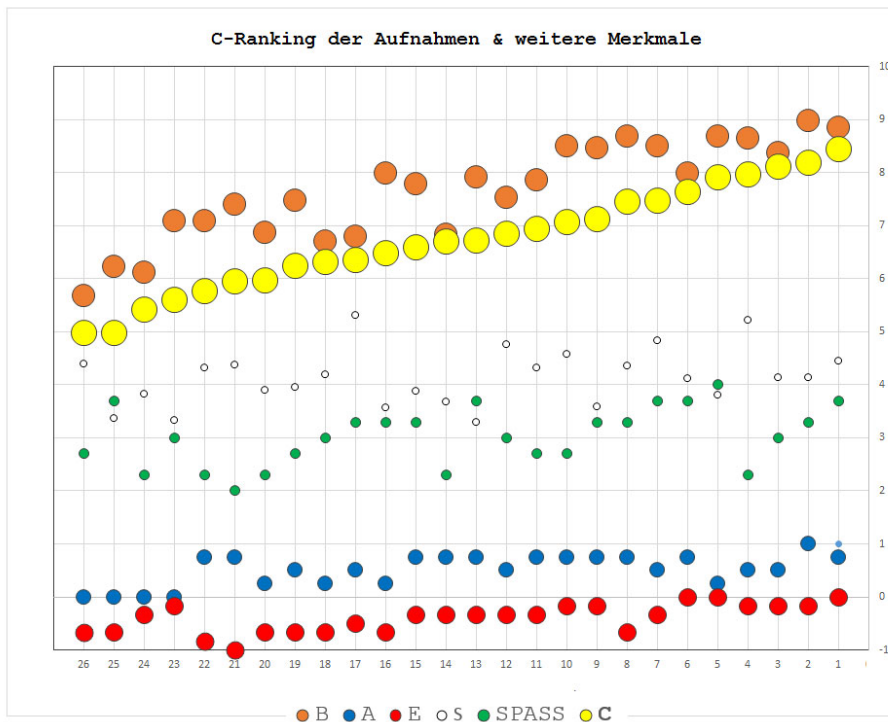
JGST fällt vor allem dadurch auf, dass kaum eine Korrelation mit C vorliegt. Dies kann darauf zurückgeführt werden, dass so viele Aufnahmen aus 6 und 7 vorliegen. Weiter fällt JGST einzig durch seine Korrelation mit S auf. Um diese Tendenz zu quantifizieren, sollen hier die JGST-spezifischen S-Mittelwerte präsentiert werden, als Referenzpunkte für Anschlussuntersuchungen mit höheren und gleichmäßigeren Schüler*innen-Zahlen (s. Tabelle 5). Für die weiterführende Analyse kann JGST jedoch als eindeutig schwächster Faktor ausgeschlossen werden. Zur Frage nach dem Eigentempo bei fortgeschrittenen Schüler*innen kann als grobe Orientierungslinie die Mittelwertreihe für S nach JGST festgehalten werden:

JGST	3 (n = 1)	4 (n = 3)	5 (n = 3)	6 (n = 10)	7 (n = 9)
S	3,28	3,73	3,91	4,08	4,37
WPM (S*60)/2	98	112	117	122	131

Tab. 5: Jahrgangsstufen-Mittelwerte auf Basis der n = 26 Aufnahmen in S (Silben pro Sekunde) und Doppsilben pro Minute (hypothet. BPM bzw. WPM).

Abbildung 7 zeigt die Kennwerte der einzelnen Aufnahmen. Die bunte vertikale „Perlenkonstellation“ zu einer Aufnahme repräsentiert das differenzierte Vorlese-Profil eines Kindes zu einem bestimmten Zeitpunkt, diagnostiziert anhand einer Prima-vista-Aufnahme des Textes „Dolmetscher“ von ca. einer halben Minute Dauer.

Explorative Aufschlüsse zum Zusammenwirken der Faktoren sollen nun über lineare Regressionsanalysen gewonnen werden. Der Faktor SPASS erwies sich dabei als schwächster Faktor, weshalb er für die folgenden Analysen ausgeschlossen wurde. Vor diesem Hintergrund soll die bisherige Evidenz noch einmal zusammengefasst werden. Ausgangspunkt dafür ist die zentrale Hypothese (s. Einführung), dass sich mit den erhobenen Daten ein Modell nach dem Muster $C \sim B + A + \text{andere Faktoren}$ bestätigen lässt. Was die Rangfolge der Faktoren angeht, korreliert B stets so stark mit C, dass der Verdacht gerechtfertigt ist, dass mit B in hohem Maße dasselbe gemessen worden ist wie mit C (Trennschärfe). Hinsichtlich der verbleibenden Faktoren sind unterschiedliche Rangfolgen festzustellen gewesen.



Anmerkungen: Rating-Mittelwerte C = „Vorlesewettbewerb“, B = „Stimmführung“, A = „Hervorhebung“, E = - Fehlersumme („Errors“), S = Silben/sek („Speed“), SPASS („Engagement“, s. Sappok, Linnemann & Stephany 2020).

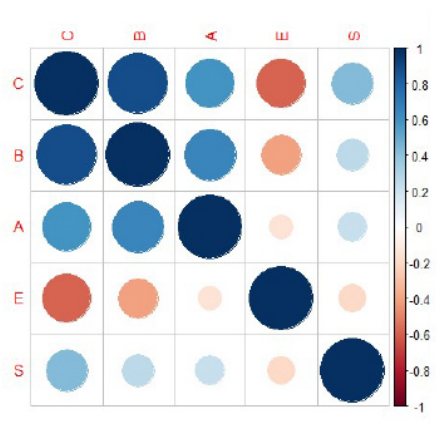
Abb. 7: Aufnahmespezifische „Advanced prosody“-Register für die Diagnostik von Vorlesekompetenz. Erfasste Merkmale zu n = 26 Aufnahmen fortgeschrittener Schüler*innen: Die Aussagekraft der Parameter im Spiegel von r-Werten ist durch die Perlengröße angedeutet.

Die A-Daten haben sich bzgl. Phrasen 4 und 5 als zu wenig reliabel gezeigt und wurden dort durch Mittelwerte ersetzt. Um das Zusammenwirken der Faktoren und hier vor allem die Rolle von Faktor A ausschließlich anhand vollständiger Datensätze zu klären, wurden diese Phrasen für die Erstellung von linearen Modellen ganz ausgeschlossen. So wurden die verbleibenden vier Audiophrasen (1, 2-1, 2-2 und 3, s. Tabelle 2) zu Aufnahme-Mittelwerten zusammengefasst. Diese repräsentieren die via C diagnostizierte Vorlesekompetenz und ihre potenziellen Komponenten. Der Unterschied schlägt sich in leichten Abweichungen der Kennwerte im Vergleich von Abbildung 6 b) und 8 a) nieder.

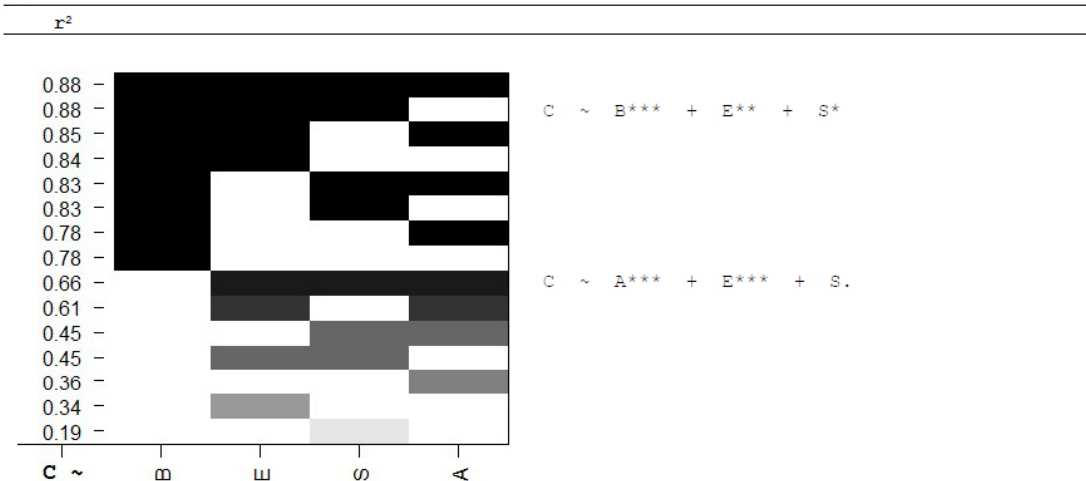
Im Hinblick auf die Korrelationsverhältnisse zeigt sich nun die Rangfolge (Abbildung 8 a): $B > A = E > S$.

a) Korrelationsmatrix Aufnahmen

	C	B	A	E	S
C	1				
B	.88	1			
A	.60	.65	1		
E	-.58	-.40	-.15	1	
S	.44	.26	.22	-.19	1



b) lineare Modelle



Anmerkungen: Das Ranking bezieht sich auf die leaps-Analyse (Lumley 2020) geordnet nach r^2 ; *** = $p < .001$, ** = $p < .01$, * = $p < .05$, . = $p < .1$ (s. Tabelle 6)

Abb. 8: Ebene der Aufnahmen ohne Phrasen 4 u. 5:

- a) Korrelationskoeffizienten r und Visualisierung (vgl. Abbildung 6).
- b) Ranking linearer Modelle zur Aufklärung von Varianz in C. Oben mit allen Faktoren, unten ohne Faktor B. Mit B + E + S und A + E + S ergeben sich zwei Modelle, wodurch die Rollen von Faktor B und Faktor A vergleichbar werden (s. Tabelle 4).

Die Exploration besteht in der Generierung und Gewichtung aller linearen Regressionsanalysen zur Prädiktion von C, die mit den Faktoren B, E, S und A möglich sind. Verwendet wurde hierfür das „leaps“ Package in R (Lumley 2020). Das Ergebnis besteht in einem Ranking der Faktorkombinationen nach in C aufgeklärter Varianz anhand von r^2 .

Zusammenfassend sollen zunächst alle Modelle betrachtet werden, die unter Beteiligung von B möglich sind (= „Pistolengriff“-Block in Abbildung 8 b) links oben). Es zeigt sich die Rangfolge $B > E > S > A$.

Der untere Teil von Abbildung 8 b) bezieht sich auf die Modelle ohne B. Hier zeigt sich eine Teilumkehrung der Ränge der verbleibenden Faktoren: $A = E > S$.

Die angebotenen Modelle wurden mit der Funktion $lm()$ in R einzeln neu gerechnet und bestätigt. Angegeben sind die zwei Modelle, die als maximal aufschlussreich erachtet werden. Beim ersten dieser Modelle (8 b) oben) wurde Faktor A ausgeschlossen, da er keinen Beitrag leistet ($C \sim B + E + S$). Beim zweiten Modell wurde Faktor B durch Faktor A ersetzt ($C \sim A + E + S$).

Bei Modell 1 ist der adj. r^2 -Wert von .86 noch einmal höher als der für das einfache Modell $C \sim B^{***}$, wo adj. r^2 bei .77 liegt. Dies spricht gegen einen kategorischen Mangel an Trennschärfe zwischen C und B. Die damit angesprochene Problematik wird in der Diskussion aufgegriffen.

Nur aus der Perspektive der beiden Modelle lässt sich eine aufschlussreiche Gegenüberstellung von Faktor B und Faktor A erreichen. Faktor A und B korrelieren nach Abbildung 8 a) mit $r = .65$ (adj $r^2 = .40$, $p < 0,001$). Vor diesem Hintergrund könnte A als eine Art „schlechter Ersatz“ für B angesehen werden. Auch dieser Befund wird in der Diskussion aufgegriffen. Ein summativer Einfluss im Sinne von $C \sim B + A$, wie er im Fujisaki-Modell (Abbildung 2 links) mit dem Prinzip Superimposition vorausgesetzt wird, konnte hier nicht belegt werden. Auch dies wird in der Diskussion aufgegriffen.

5 | Diskussion und Ausblick

Untersucht wurde das diagnostische Potenzial von 26 Audioaufnahmen eines Sachtextes von ca. 30 Sekunden Dauer. Es handelte sich um genau diejenigen Aufnahmen aus einer Längsschnitt-Stichprobe von $N = 99$, die auf der NAEP-Fluency-Skala einhellig Level 4 zugeordnet bekamen (Sappok, Linnemann & Stephany 2020). Jede Aufnahme wurde in 6 Phrasen von max. 22 Textsilben Länge geschnitten und es ergab sich eine Stichprobe von 156 Audiophrasen, auf der die Untersuchung beruht. Hierfür wurde eine dreizügige Ratingprozedur entwickelt und mit einer Stichprobe von 39 studentischen Rater*innen im Rahmen eines E-Seminars erprobt. Alle Rater*innen bewerteten alle 156 Audiophrasen am Computer als Hörstimuli in individuell teilrandomisierter Form. Team A bewertete informationsstrukturell motivierte phonetische Wortprominenz auf der binären Skala A („Zielwort hervorgehoben vs. Gegenbeispiel“). Team B benotete die Phrasen-Intonationskontur auf der vierstufigen Skala B („Stimmführung insgesamt satzzeichengemäß?“) mit den Antwortmöglichkeiten: sehr gut, gut, befriedigend, ausreichend oder weniger. Team C benotete Vorlesekompetenz global durch die Vergabe von 1 bis 10 Punkten in einem Casting-Szenario.

In der vorliegenden Untersuchung wurde somit weiterführende Grundlagenforschung betrieben, deren Befunde nun im Hinblick auf ihre Perspektiven diskutiert werden sollen. Zu nennen ist hier einmal die Perspektive, wie die Grundlagenforschung selber weitergeführt wird (z. B. Optimierung von Skalenkonstruktion). Hierzu wurden an den Anfang zwei Heurismen gestellt. Diese sollen nun im Unterkapitel Diagnostik diskutiert werden. Außerdem ist die Perspektive zu diskutieren, wie die in vorliegender Untersuchung generierte Evidenz zur Praxis weitergeführt werden kann. Dazu wurden konkretere Hypothesen aufgestellt. Diese Hypothesen sollen im Unterkapitel Didaktik aufgegriffen werden. Zum Bezug zum Textverstehen wurden zwei spekulative Axiome formuliert. Auch wenn der Bezug zum Textverstehen in vorliegender Untersuchung nicht im Zentrum stand, sollen diese Axiome in der Zusammenfassung noch einmal angesprochen werden.

5.1 | Diagnostik

Zunächst wird die Weiterarbeit aus Grundlagenperspektive fokussiert. Die Heuristiken I und II haben sich grundsätzlich bestätigt. Als Ausgangspunkt für die konkrete Weiterarbeit lassen sich die folgenden Anknüpfungspunkte für eine Optimierung des behandelten Diagnostikansatzes definieren:

- bessere theoretische Fundierung (Konstruktebene)
- besserer Lesetext (Auswahl oder Konstruktion von Textmaterial für die Diagnostik)
- bessere diagnostische Instrumente (Skalenkonstruktion, Benutzeroberfläche)
- „bessere“ Rater*innen (Expert*innen, z. B. aufwändiger geschulte Studierende oder Lehrkräfte)

Bei der Untersuchung der Inter-Rater-Reliabilität zeigte sich klar $C \gg B > A$. Was den DOL-Text angeht, so können die Phrasen 1 bis 4 als geeignet für die Erhebung von A gelten, Phrasen 4 und 5 nicht. Als Ansatz für eine Verbesserung der Skala kann gelten, dass die Hörbedingungen verbessert werden müssen. Als Orientierung kann dabei weiterhin die Akzent-Komponente nach Fujisaki (1988, S. Abb. 3) gelten.

Auch Skala B ist verbesserungsbedürftig. Auf Konstruktebene ist mit dem Bezug zur Zeichensetzung und damit den grobsyntaktischen intonationsrelevanten Verhältnissen ein prosodischer Aspekt angesprochen, der gegenüber A weniger gehobene Aspekte von Textverstehen voraussetzt. Die untersuchten fortgeschrittenen Schüler*innen lassen in puncto syntaxgemäßer Stimmführung häufig nichts zu wünschen übrig. Es hat aber den untersuchten Rater*innen womöglich gewisse Schwierigkeiten bereitet, B zu fokussieren und andere Eigenschaften von gutem Vorlesen dabei auszublenden (die Korrelation $B \sim C$ liegt bei $r = .9$, vgl. Abb. 6 b).

In Hinblick auf diese Probleme sind bzgl. zukünftiger Rater*innen-Auswahl verschiedene Maßnahmen zu erwägen, z. B. gratifizierte Schulungsmaßnahmen mit Beispiel-Audiophrasen, Rekrutierung aus anderen Populationen (z. B. Lehrkräfte). Außerdem sollte bzgl. der Skalen A und B über Modifikationen der Konzeption und der Instruktion in den Skalen selbst nachgedacht werden. Demgegenüber kann aber festgehalten werden, dass es mit Skala C offenbar gelingt, ein latentes Personenmerkmal "Vorlesekompetenz" reliabel zu erheben.

5.2 | Didaktik

Abschließend wird vor diesem Hintergrund die Weiterarbeit aus Praxisperspektive fokussiert. Im Mittelpunkt des vorliegenden Beitrags stand folgende

zentrale Hypothese: Die erhobenen Ratingdaten können zu einem explorativen Modell nach dem Muster $C \sim B + A + \text{weitere Faktoren}$ integriert und methodisch und didaktisch weiterführend interpretiert werden.

Der Befund, dass im schlimmsten Fall mit B etwas sehr Ähnliches wie mit C gemessen wird, ist ernüchternd und zeigt auf, dass Skala B konzeptionell überarbeitet werden muss. Der Befund, dass mit A etwas Ähnliches wie mit C gemessen wird, ist hingegen in unerwarteter Weise aufschlussreich. Dass ein minimales Kontingent von bereits vier binären Items genügt, um Aspekte

von Vorlesekompetenz in nennenswertem Maße widerzuspiegeln, zeigt, dass Akzent- bzw. Prominenzverhältnisse mit verbesserten Methoden weiter untersucht werden sollten, zumal hier Textverständnis recht eindeutig als Vorbedingung angesehen werden kann. Hervorzuheben ist, dass die Analyse von Hervorhebung auf Wortebene oder ein vergleichbares, optimiertes Konstrukt mit signalphonetischen Mitteln erfolgen könnte, da es sich eher als andere Merkmale automatisch lokalisieren lässt.

Für die Gültigkeit der Hypothesen, die zum Abschluss von Unterkapitel 2 aufgeführt wurden, Hypothese 1: Aus 50% aller Grundschüler*innen werden bis Mitte Sekundarstufe I keine flüssigen Leser*innen,

Hypothese 2 a) Das Überschreiten der Schwelle zu NAEP-Level 4 bedeutet, lediglich early prosody erworben zu haben,

Hypothese 2 b) Innerhalb von NAEP-Level 4 (advanced prosody) ist ein breites Spektrum von Binnenabstufung auszudifferenzieren,

konnten bereits in Sappok & Fay (2018) und Sappok, Linnemann & Stephany (2020) Hinweise gesammelt werden. Hinsichtlich Hypothese 1 lassen die geringen Stichproben allerdings keine Generalisierung zu. Zu Hypothese 2 a) und b) kann hingegen verallgemeinert werden, dass die Pinnell-Skala sehr anfällig für Deckeneffekte ist (ebd.). Hier setzen die untersuchten Skalen an. Skala C kann als voll einsatzfähig für die Diagnose von advanced prosody gelten. Dies bedeutet jedoch eine Abkehr vom Begriff Leseflüssigkeit im engeren Sinne, denn es spielen auch Faktoren wie Stimme-Verstellen eine Rolle, während Geschwindigkeit kaum mehr eine Rolle spielt (s. u.). Die Skalen B und A können aus unterschiedlichen Gründen noch nicht als einsatzfähig gelten. Die Untersuchung hat aber konkrete Anknüpfungspunkte für die Weiterarbeit ergeben.

Modell	adj. R ²	Faktor	b	SE	t	df	p
1	.86	B	0,96	0,11	-8,89	22	< .0001
		E	-0,77	0,25	-3,14	22	.0048
		S	0,46	0,18	2,57	22	.0173
2	.62	A	1,79	0,48	3,73	22	.0012
		E	-1,43	0,39	-3,70	22	.0013
		S	0,57	0,30	1,89	22	.0727 (n. s.)

Anmerkungen: abhängige Variable: C = Vorlesewettbewerb, unabhängige Variablen: B = Stimmführung, E = Errors, S = Speed, A = Hervorhebung

Tab. 6: Lineare Modelle zur Gegenüberstellung von Faktor A und B als Komponenten von C (s. Abb. 9). Geschwindigkeit (S) spielt die geringste Rolle.

Aus schulpraktischer Perspektive steht die Frage im Vordergrund, wie didaktische Ressourcen optimal investiert werden können. Dazu lassen sich in einem letzten Schritt die in den Modellen aufgezeigten Steigungsverhältnisse auswerten (Spalte b in Tabelle 6). Dabei werden fünf „Advanced-Prosody-Faustregeln“ formuliert, die für den Unterricht Orientierung bieten könnten. Modell 1 (Spalte b) ist so zu lesen: Wenn jew. die beiden anderen Faktoren konstant bleiben,

gelten folgende Gesetzmäßigkeiten: für + 1 B (hochskaliert auf 1 bis 10) gibt es + 0,96 C-Punkte; für + 1 Error pro Phrase gibt es – 0,77 C-Punkte; für + 1 Silbe pro Sekunde gibt es + 0,46 C-Punkte.

Der Befund, dass B und C nicht trennscharf genug sind, tut der potenziellen Relevanz von Stimmführung keinen Abbruch. Für die didaktische Arbeit kann hierzu folgende Faustregel eingesetzt werden:

(1) *Beachte deine Stimmführung auf dem Weg zum Satzzeichen.*

Auch Faktor E spielt eine nennenswerte Rolle. Hinsichtlich der Fehler kann wie in der Grundschule nach wie vor die Faustregel ausgegeben werden:

(2) *Mache keine Flüchtigkeitsfehler, finde deine Problemfelder.*

Der Einfluss von S, einem der wichtigsten Faktoren bei der Diagnose von Leseflüssigkeit im Primärkontext, ist gering. Didaktischer Handlungsbedarf in Hinblick auf eine Steigerung besteht höchstens im Andante-Bereich (bis $S = 3,6$ bzw. 108 BPM). Die Steigung ist bei S vordergründig dahingehend zu interpretieren, dass + 1 S fast so viele C-Pluspunkte bringt wie - 1 E, also ein Fehler weniger pro Phrase. Hält man sich vor Augen, dass + 1S etwa + 30 BPM entspricht, erweist sich eine Investition in Geschwindigkeit im Spiegel des Modells allerdings klar als uninteressant. Wahrscheinlich profitieren manche Schüler*innen eher von einer beschwichtigenden Faustregel:

(3) *Finde dein Eigentempo. So schnell wie möglich zu lesen war nur in der Grundschule wichtig.*

Die Steigungsbefunde des zweiten Modells (Tabelle 6) können folgendermaßen interpretiert werden: Wenn jew. die beiden anderen Faktoren konstant bleiben, gelten folgende Gesetzmäßigkeiten: für + 1 A gibt es + 1,79 C-Punkte; für + 1 Error pro Phrase gibt es – 1,43 C-Punkte; für + 1 Silbe pro Sekunde gibt es + 0,57 C-Punkte.

Die Relevanzverhältnisse lassen sich bzgl. E und S recht gut mit denen zu Modell 1 vergleichen. Für A kann aus didaktischer Perspektive festgehalten werden, dass mit knapp 2 Pluspunkten ein hoher C-Profit winkt, wenn man lernt, einer informationsstrukturell gebotenen Hervorhebung Rechnung zu tragen. Als Faustregel lässt sich das z. B. so formulieren:

(4) *Wörter, auf die es ankommt: erkenne und betone sie.*

5.3 | Zusammenfassung

Eine Zusammenfassung zur vorliegenden Untersuchung sollte mit dem Begriff *advanced prosody* beginnen. Dieser Begriff wurde gewählt, um die Notwendigkeit einer Ausdifferenzierung des Konstrukts Leseflüssigkeit hervorzuheben. Die herkömmlichen Verwendungskontexte von *reading fluency* beziehen sich auch international fast ausschließlich auf den Primärbereich. *Advanced prosody* in der weiterführenden Schule zeichnet sich vor allem dadurch aus, dass Geschwindigkeit kaum noch relevant ist. Vor diesem Hintergrund muss der Begriff *reading fluency* als *genus proximum* für *reading prosody* hinterfragt werden, denn für Forschung und Praxis gehen mit *advanced prosody* andere Probleme einher als mit *early prosody*. Diese Probleme wurden mit vorliegender Untersuchung vor dem Hintergrund folgender „spekulativer Axiome“ exploriert (siehe Einführung):

Axiom I: Kompetenz beim lauten Lesen ist ein Indikator für Aspekte von Textverstehen.

Axiom II: Umgekehrt können solche Aspekte über die Förderung von lautem Lesen mitgefördert werden („bootstrapping“).

Hinsichtlich Axiom I haben sich trotz verbleibender Desiderate neue Hinweise auf die Relevanz von Informationsstruktur als Indikator für Textverständnis auf hohem Niveau und damit diagnostisches Potenzial ergeben.

Hinsichtlich Axiom II kann zum Abschluss als didaktischer Ansatz das „Verstehen durch Sich-selber-Vorlesen“ angeregt werden. Eine solche Art der Förderung von Textverstehen über eine Förderung von lautem Lesen ist wahrscheinlich erst angebracht und erfolgreich, wenn die Gefilde von advanced fluency erreicht sind. Unterschiedlich gezielt trainiert werden könnte neben einem *publikumswirksamen* lauten Lesen demnach auch ein *epistemisches* lautes Lesen, das beim leisen Lesen „im Notfall“ zu Rate gezogen wird. Eine entsprechende Faustregel wäre:

(5) Zusammenhänge kann man hören: *Wenn du einen Satz nicht verstehst, lies ihn dir mal selber vor.*

6 | Literatur

- Beuschel-Menze, H., Ferber, M., Grimm, H., Grimm, I., Menze, F. & Wiebel, K. H. (2009). *Das große Schülerlexikon von A-Z*. cbj.
- Blaauw, E. (1995). On the perceptual classification of spontaneous and read speech. Om Dissertations Series. *Research Institute for Language and Speech*, Utrecht University.
- Boersma, P. & Weenink, D. (2020). *Praat: doing phonetics by computer* [Computer program]. Version 6.1.16. <http://www.praat.org/> [16.09.2021].
- Daane, M. C., Campbell, J. R., Grigg, W. S., Goodman, M. J., & Oranje, A. (2005). *Fourth-grade students reading aloud: NAEP 2002 special study of oral reading*. U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. Government Printing Office.
- Dahan, D. (2015). Prosody and language comprehension. *WIREs Cognitive Science*, 6(5), 441-452.
- Fisseni, B. (2011): *Focus: Interpretation? Empirical Investigations on Focus Interpretation*. Universitätsverlag Rhein-Ruhr KG an der Universität Duisburg-Essen.
- Fujisaki, H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. *Vocal physiology: Voice production, mechanisms and functions* 2, 347-355.
- Funke, R. (2018). Lautes Lesen – Was ist das und wozu dient es? *Leseräume* 5, 88-95.
- Gamer, M., Lemon, J. & Singh, I. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1. <https://CRAN.R-project.org/package=irr> [16.09.2021].
- Godde, E., Bosse, M.-L. & Bailly, G. (2020). *A review of reading prosody acquisition and development*. University Grenoble Alpes.
- Groen, M. A., Veenendaal, N. J. & Verhoeven, L. (2018). The role of prosody in reading comprehension: evidence from poor comprehenders. *Journal of Research in Reading*, 42(1), 37-57.
- Hallgren, Kevin A. (2012): Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol*, 8(1), 23-34.

- Holle, K. (2006). Flüssiges und phrasiertes Lesen (fluency). In: S. Weinhold (Hrsg.), *Schriftspracherwerb empirisch. Konzepte – Diagnostik – Entwicklung*. Schneider Verlag Hohengehren.
- Kegel, G. (1990): Sprach- und Zeitverarbeitung bei sprachauffälligen und sprachunauffälligen Kindern. In: G. Kegel u.a. (Hg.): *Sprechwissenschaft und Psycholinguistik 4. Beiträge aus Forschung und Praxis* (S. 229–255). Westdeutscher Verlag.
- KMK (2004). *Bildungsstandards im Fach Deutsch für den Primarbereich*. Beschluss vom 15.10.2004. Bonn: Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.
- Kuhn, M. R., Schwanenflugel, P. J. & Meisinger, E. B. (2010). Aligning Theory and Assessment of Reading Fluency: Automaticity, Prosody, and Definitions of Fluency. *Reading Research Quarterly*, 45, 230-251.
- Laver, J. (1994). *Principles of Phonetics*. Cambridge University Press.
- Lenhard, W. & Schneider, W. (2006). *ELFE 1-6. Ein Leseverständnistest für Erst- bis Sechstklässler*. Hogrefe.
- Lenhard, W. & Lenhard, A. (2014-2017). *Berechnung des Lesbarkeitsindex LIX nach Björnson*. <http://www.psychometrica.de/lix.html> [16.09.2021]. Psychometrica.
- Lumley T. (based on Fortran code by Alan Miller) (2020). *leaps: Regression Subset Selection. R package version 3.1*. <https://CRAN.R-project.org/package=leaps> [16.09.2021].
- Miller, G. A. (1956): The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review* 63, 81–97.
- Möbius, B. (1993). *Ein quantitatives Modell der deutschen Intonation. Analyse und Synthese von Grundfrequenzverläufen*. Narr.
- Nix, D. (2011). *Förderung von Leseflüssigkeit. Theoretische Fundierung und empirische Überprüfung eines kooperativen Lautlese-Verfahrens im Deutschunterricht*. Juventa.
- Pinnell, G. S., Pikulski, J. J., Wixson, K. K., Campbell, J. R., Gough, P. B. & Beatty, A. S. (1995). *Listening to children read aloud. Data from NAEP's Integrated Reading Performance Record (IRPR) at Grade 4*. Office of Educational Research and Improvement, U.S. Department of Education.
- Pompino-Marschall, B. (2009). *Einführung in die Phonetik*. De Gruyter.
- Pöppel, E. (1997): A hierarchical model of temporal perception. *Trends in Cognitive Sciences* 2, 56–61.
- R Core Team (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. <http://www.R-project.org/> [16.09.2021].
- Rietveld, T. & Chen, A. (2006). How to Obtain and Process Perceptual Judgments of Intonational Meaning. In: S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter & J. Schließer (Hg.): *Methods in Empirical Prosody Research* (S. 283–320). De Gruyter.
- Rosebrock, C. & Nix, D. (2006). Forschungsüberblick: Leseflüssigkeit (Fluency) in der amerikanischen Leseforschung und -didaktik. *Didaktik Deutsch*, 20, 90–112.
- Rosebrock, C. & Nix, D. (2015). *Grundlagen der Lesedidaktik und der systematischen Leseförderung*. Schneider.
- Rosebrock, C., Nix, D., Rieckmann, C. & Gold, A. (2016). *Leseflüssigkeit fördern. Lautleseverfahren für die Primar- und Sekundarstufe*. Kallmeyer.
- Röttig, S., Schwerkolt, C. & Nottbusch, G. (2021). Die Entwicklung der Leseflüssigkeit in der Grundschule. Eine Longitudinalstudie über die interagierenden Dimensionen Dekodiergenauigkeit, Automatisierung, Lesegeschwindigkeit und Prosodie bei Kindern der Jahrgangsstufen 2 und 3 (in diesem Band).

- Sappok, C. (2002). *Metrum, Füllung und Signal. Eine Untersuchung der Betonungsintervallldauern bei 32 Aufnahmen einer Strophe aus dem Gedicht „Bimini“ von Heinrich Heine*. Magisterarbeit, Technische Universität Berlin.
- Sappok, C. (2010): The quantitative organization of speech. *Proc. Speech Prosody 2010*, paper 102
- Sappok, C. & Arnold, D. (2012a): On the Normalization of Syllable Prominence Ratings. In: *Proceedings of Speech Prosody, 6th International Conference of the International Speech Communication Association*.
- Sappok, C. & Arnold, D. (2012b): More on the Normalization of Syllable Prominence Ratings. In: *Proceedings of Interspeech, 13th Annual Conference of the International Speech Communication Association*.
- Sappok, C. & Fay, J. (2018): Prosodische Aspekte von Leseflüssigkeit messen. Evaluation einer Rating-prozedur mit Audioaufnahmen von DrittklässlerInnen. *Didaktik Deutsch*, 44, 61-83.
- Sappok, C., Linnemann, M. & Stephany, S. (2020): Leseflüssigkeit – Prosodie – Leseverstehen. Eine Longitudinalstudie zur Entwicklung der Leseflüssigkeit von Jahrgangstufe 3 bis 7. In: I. Rautenberg (Hrsg.): *Evidenzbasierte Forschung zum Schriftspracherwerb* (S. 175-209). Schneider Verlag Hohengehren.
- Schwanenflugel, P. J., Westmoreland, M. R. & Benjamin, R. G. (2013): Reading fluency skill and the prosodic marking of linguistic focus. *Reading and Writing*, 26(6), 9-30.
- Stephany, S., Linnemann, M., Goltsev, E. & Becker-Mrotzek, M. (2021). Prosodische Aspekte der Leseflüssigkeit als Indikator für Lesekompetenz. Analysen mithilfe prosodischer Lupenstellen (in diesem Band).
- Tillmann H. G. & Mansell, P. (1980): *Phonetik*. Klett-Cotta.
- Weaver, W. & Shannon, C. E. (1949): *The Mathematical Theory of Communication*. Urbana Ill.
- Wimmer, H. & Mayringer, H. (2014). *Salzburger Lese-Screening für die Schulstufen 2-9*. Hogrefe.