Oral reading proficiency and prosody – a perceptual pilot study on especially fluent German students (grade 3 to 7)

Christopher Sappok

Institute of German Language and Literature II, University of Cologne, Germany
csappok@uni-koeln.de

ABSTRACT

Phonetic research on children's read speech is rare compared to educational research. In this pilot study, the established educational NAEP-scale [1] for perceptual ratings of reading fluency is juxtaposed to several phonetic, mainly prosodic, parameters. From a greater corpus we focus on recordings in which performance was identified as maximally fluent on the NAEP scale by university student raters with maximal reliability, arriving at a subsample of 50 recordings of 15 subjects reading two texts. We found a considerable amount of systematic variance in a number of our partly newly introduced perception- and measure-based variables, which are thus relevant for quantifying the developmental stage of *oral reading proficiency*, establishing a more advanced notion of fluency. (Note that all variance found is evidence for a ceiling effect of the NAEP fluency scale.) The findings pave the way for a phonetically sound target modelling of oral reading proficiency development.

Keywords: oral reading, read speech, fluency, prosody, German children

1. INTRODUCTION

In our context, several lines of research need to be distinguished, namely phonetic research on natural speech and on read speech and educational research on reading acquisition. In phonetics, one often has to rely on written stimuli in order to control for content, even when the eventual objective is maximally natural speech. Read speech as an objective in its own right is underrepresented. Earlier examples are [2] on Swedish and [3] on Dutch grown up readers. An early example on children's read speech is [4] (American English), who see prosody as a *space* and read speech as a *variety*: "The target is in effect to define a prosodic space, with dimensions corresponding to the relevant variables, and different varieties occupying different regions of the space" (other varieties mentioned in [4]: expression of emotion, functional distinctions, social varieties, and clinical conditions). In order to arrive at such an (intersubjective or even interlingual) space we need to understand intrasubjective relations between read and spontaneous speech as well as within subject reading skill development. Our approach focuses on development. An approach focusing read vs. spontaneous speech with German children is [5], who points out the problem of the target model relevant with all efforts to evaluate developmental stage: "child language researchers interested in phonetic and phonological acquisition need to be very careful when collecting adult speech data to determine the target model" [5]. This issue leads to educational research, based on the term *fluency*. Two recent meta analytic studies [6, 7] give an overview, together covering 43 studies from 1995 to 2019. Even though both meta studies have the term *prosody* in their titles, a majority of the included studies does not fully follow phonetic standards on the matter, e. g., in [8] a pitch target model for evaluating early elementary school children is coarsely calculated from measures in adult readings of the same text; in [9] the target model for reading acquisition is coarsely equated with spontaneous speech. Further, most studies are limited to elementary school contexts, leaving blank the "teenage" period between elementary and adult oral reading. On the other hand, this line provides a definition based, well established rating procedure. An agreed upon fluency definition is subsumed in [9], naming the following dimensions: *accuracy*, *automaticity*, and *oral reading prosody*. The underlying concepts are mirrored in the following instruction for listeners to detect fully developed fluency used by the National Center for Education Statistics (NAEP) of the United States: "Reads primarily in larger, meaningful phrase groups. Although some regressions, repetitions, and deviations from text may be present, these do not appear to detract from the overall structure of the story. Preservation of the author's syntax is consistent. Some or most of the story is read with expressive interpretation" [1]. With its mixture of relevant but very different facets this formulation represents level 4 on the 4-level-NAEP fluency scale. In our recent work we found strong hints that this scale produces a pronounced ceiling effect, when applied (in translated form) to longitudinal audio stimuli from grade 3 to 7 [10, 11]. To quantify the extent of this ceiling effect, the present study investigates mainly recordings identified as NAEP-level 4, and uses alternative hypothetically relevant parameters to disentangle the very different features mentioned "in one go" in the level-description above. We will present arguments for the parameters' relevance as well as evidence from descriptive and correlational analysis. Our main research question is concerned with perception: Which rating scales reliably capture oral reading proficiency? Only when this is sufficiently clear we are ready to fully tackle the next question: Which relations may be established between the according perceptual parameters and parameters that can be measured in the signal? [12]

2. SAMPLE

2.1. The DOL-recordings of the LAUDIO-corpus

The LAUDIO-corpus (= Longitudinal AUDIO) is a collection of audio files from German pupils reading aloud. The underlying subject sample consists of 4 German elementary school classrooms (2-times 3rd grade, 2-times 4th grade, $N_{LAUDIO} = 61$), all of the same "Inklusive Grundschule" (desired standard in Germany), in a rural part of the Federal State of Rheinland-Pfalz. The design was to record all children and then again once a year, 4 recording rounds altogether, the biographical windows covered being grade 3 to 6 for the original 3rd graders and grade 4 to 7 for the original 4th graders. Rounds were held in winter, the first being winter 2014/2015, the last 2017/2018. (Note that thus these windows are almost the very last windows that were totally unaffected by the Covid pandemic.) In each round, children were recorded in sessions of roughly 15 minutes on being asked to read aloud to the experimenter (author) a variety of text stimuli, each prima vista and secunda vista. Two anchor texts were used in every session from the first round on: DOL, an entry of 53 words on the lemma "Dolmetscher" (interpreter) in a youth's lexicon, estimated grade level: 6; and SCHNEE, a fable-like story ("The hare and the snowman"), 191 words, mainly dialogue, estimated grade level: 3. The validity of the subject sample is limited by factors such as school and school type, state, city-country-gap etc., and attrition.

2.2. Subsample selection

Many of the LAUDIO recordings are too poor to be included in a setting suitable for the perceptive isolation of distinct prosodic parameters. These were identified in a screening procedure involving 247 recordings divided into 17 subsets á 14 or 15 files and 51 university student raters (BA and MA in teaching contexts, but non expert in phonetics). Each rater was randomly placed at a headphone plus computer station displaying a spreadsheet plus one of the audio subsets and given a paper including the NAEP scale. After a 15-minute introductory session, raters self activated the files repeatedly and noted their NAEP level evaluations (further details in [10]) in the spreadsheet. This way each subset was rated by three raters. Inter rater reliability was computed via intra class correlation (ICC) analysis using the R-package *irr* ([13], model = twoway, type = consistency, unit = average). The median of the subsets' ICC coefficients was > .8 ("excellent" [14], details in [10]). Now, based on the resulting rating means per recording, two subsamples DOL+ (fluent) and DOL++ (especially fluent) were selected using only prima vista recordings. For DOL+ we eliminated subjects with unsuccessful biographies, i.e. those of which no recording with NAEP fluency > 3 is available at all, arriving at 144 recordings from $N_{DOL+} = 31$ subjects. Further, for DOL++ all recordings were eliminated that had not been rated NAEP fluency = 4 by all three raters (*post hoc* ICC coefficient = 1). This left us with 26 recordings from $N_{DOL++} = 16$ subjects. Since all recordings are specified with respect to subject, year and vista, there is a SCHNEE-twin for almost every one of the DOL recordings. Based on the selection process, we assumed these twin SCHNEE recordings to represent equivalent developmental stage and included them in the analyses presented here. For economic reasons, with the SCHNEE recordings we used only the sequence of the first 120 of its 191 words in perception and measure analysis.

3. METHODS

3.1. Dependent variable: overall proficiency [C]

There is no rating scale for advanced fluency as established as the NAEP fluency scale for earlier fluency. In our approach to this *desideratum* we waived explicating any hypothetically relevant prosodic features because there is so little research on post elementary oral reading. Posing specific listening citeria would have been not hypothetical but speculative. Instead we developed a holistic scenario of the rater being part of a jury of a fictious school's oral reading championship, confronted with a 10-point scale as commonly known from popular contexts such as figure skating. The only hint was that the championship was *omnium contra omnem* to keep the raters from integrating perceived voice age in their ratings ("pardon young effect"). The championship scale [C] was introduced and successfully evaluated in [11]. Due to the largeness of the + samples, the C scale was applied to the DOL+ sample and the SCHNEE+ sample with only three fixed expert raters (including the author). This time the rating session was conducted in a fully automated computer setup prepared with the ExperimentMFC functions (including randomisation of audio stimulus presentation) provided by praat [15]. ICC coefficients were > .8 with the DOL+ and > .7

with SCHNEE+ sample ([13], model as above, smaller reliability with SCHNEE can be attributed to specifics of the text).

3.2. Basic independent variables: deviation from text [SMS, E] and duration [T]

We count all text deviating sequences and obviously unplanned pauses > 1s (otherwise regardless of actual length) as occurrences of special mixup sequences [SMS]. In these cases readers are quite obviously confused (or “mixed up”) and cognitively busy regaining control. Often, longer duration of such a sequence represents thoroughness of self monitoring rather than lack of proficiency. As errors proper [E] we count deviations on the word level (i.e. deviation sequences < 1s, otherwise regardless of actual length and of self correction). With respect to duration, the measures established in educational research are words/minute (too coarse on our opinion) or correct words/min (even coarser, because this mixes the dimensions accuracy and automaticity mentioned above [9]). In our context we capture the latent trait of habitual tempo (details see [11]) by erasing from the signal all SMS sequences and all inter phrasal pauses regardless of length and dividing the result’s duration by the number of speech syllables, arriving at speech rate or habitual tempo T [syllables/s].

3.3. Perceptual independent variables: phrasing [P] and rhythm [R]

There is no room here to discuss phrasing and rhythm sophisticatedly from a phonetic perspective. From an educational perspective it is important to note that prosodic concepts should be comprehendable for children. Otherwise it would be impossible to explain to them what they should work on when according deficits are detected in their performance. Based on this consideration we developed a dichotomous specification of “expressiveness” [1]: clarity of phrasing [P] and diversity of rhythm [R], with rhythm subsuming overall liveliness and phrase level accentuation. For both P and R, 4-level Likert scales were applied to the DOL+ sample and the SCHNEE+ sample in the same ExperimentMFC based expert rater setup as with the C scale (see above). With DOL+, ICC coefficients were > .8 for P and R, with SCHNEE+, ICC coefficients were > .7 for P and R.

4. RESULTS

4.1. The DOL+ subsample

The present paper is focused on especially fluent children, i.e. the ++ subsamples, but note that ++ is a subset of +, meaning that the ++ perceptual values for C, P and R were obtained in the according + reference fields. Within such listening reference fields, certain self normalization tendencies need to be taken into account (see [16] for our work on the normalization of rating data). Tab. 1 gives a descriptive overview of DOL+ as it is the reference field most relevant here.

	mean	sd	min	max
NAEP	3.70	0.33	3.00	4.00
C	7.34	1.10	4.00	9.67
P	2.81	0.64	1.00	4.00
R	2.94	0.67	1.33	4.00
E	3.36	2.26	0.00	10.00
SMS	0.30	0.63	0.00	3.00
T	4.05	0.73	2.31	5.79
AGE	11.17	1.38	8.24	13.74
GRADE	5.21	1.25	3.00	7.00

Tab. 1: Descriptive overview of the DOL+ subsample (144 recordings of 31 subjects reading a quite difficult text of 53 words). C represents overall oral reading proficiency.

NAEP mean and sd indicate skewedness towards 4 which is first evidence of a ceiling effect of this scale when applied to post elementary school contexts (see GRADE). C values display that raters rather make use of the upper half of the 10-point C scale when confronted with recordings from the upper third of the NAEP spectrum. P and R appear to behave in similar ways. E is high with respect to text stimulus length = 53 words, the range is broad. SMS shows the same tendency but seems less important.

4.2. The DOL++ and SCHNEE++ subsamples

	mean	sd	min	max
C	7.77	0.98	5.00	9.33
P	2.92	0.66	1.00	4.00
R	3.09	0.65	1.33	4.00
E	2.44	1.76	0.00	6.00
SMS	0.16	0.37	0.00	1.00
T	4.50	0.59	3.55	5.79
AGE	11.67	1.31	8.24	13.29
GRADE	5.84	1.14	3.00	7.00

Tab. 2: Descriptive overview of the DOL++ subsample (25 recordings of 16 subjects). Note that NAEP = 4 with all recordings.

In the first step, DOL++ description (Tab. 2) is compared with DOL+ description (Tab. 1). C mean is only slightly higher with DOL++ than it is with DOL+, showing that ++ is not so ++ after all. This points at another problem that appears to have occurred with our method of selecting DOL++: Presumably NAEP = 4 has been attributed to young readers more freely, taking account of perceived voice age, thus making the NAEP scale a referential one not only with respect to the rater's reference field but also with this "pardon young"-effect. Note that with the C scale this effect is checked with the scenario description of a fictious school holding an internal oral reading championship *omnium contra omnem.*

In the second step DOL++ description (Tab. 2) is compared to SCHNEE++ description (Tab. 3). C mean doesn't seem to take account of text difficulty. This is an argument for calculating means of the two (our best estimation for the proficiency latent trait in connection with individual developmental stage) without any normalisation [16]. Note that the better part of the SCHNEE text stimulus sequence considered consists of dialogue, prepared so that speaker identity had to be marked by the reader with prosodic means only. This "reported speech" component goes beyond the mere genre differentiation of factual text (DOL) vs. narrative text (SCHNEE), when it comes to oral reading prosody, giving a greater "playground" for prosodic variation. Accordingly the P and R means are higher with SCHNEE++, showing no self normalisation effect as with C.

	mean	sd	min	max
C	7.65	0.94	6.00	9.67
P	3.40	0.54	2.00	4.00
R	3.39	0.47	2.33	4.00
E	2.36	2.25	0.00	9.00
SMS	0.28	0.54	0.00	2.00
T	4.97	0.55	4.09	6.02
AGE	11.67	1.31	8.24	13.29
GRADE	5.84	1.14	3.00	7.00

Tab. 3: Descriptive overview of the SCHNEE++ subsample (N = 25 recordings of 16 subjects reading a quite easy text sequence of 120 words).

The E value of SCHNEE is remarkable with respect to the fact that in our operationalisation errors are simply counted and not set in proportion to either time or number of words. A comparison of, say, percentage of words done wrong, yields 4,6% with DOL vs. 2% with SCHNEE. Here text difficulty shows its face at last and the same holds for SMS and T.

In the third step means of the DOL++ and the SCHNEE++ values (except E and SMS) are calculated. The results no longer represent features of recordings but of developmental states of oral reading proficiency and its hypothetical components. With the text deviation measures E and SMS we simply sum up DOL++ and SCHNEE++ values. The outcome delivers no new information so we can directly proceed to correlation analysis.

5. DISCUSSION

The main research question finishing our introduction is: Which rating scales reliably capture oral reading proficiency? We presented adequate answers in the form of the newly introduced rating scales C, P and R. It could be shown that C well dissolves the ceiling effect of the NAEP fluency scale.

	C	P	R	E	SMS	T	AGE	GRADE
C	1							
P	.68	1						
R	.54	.72	1					
E	-.42	-.28	-.49	1				
SMS	-.48	-.45	-.48	.19	1			
T	.16	-.33	-.12	.07	.09	1		
AGE	.19	-.21	-.20	.27	.13	.60	1	
GRADE	.33	-.08	-.04	.14	.05	.59	.97	1

Tab. 4: Correlation matrix of the DOL++ and SCHNEE++ unified values (representing 25 child-timepoint combinations).

Tab. 4 shows that P and R capture distinct relevant prosodic aspects of oral reading that contribute a considerable deal to C. Deviation from text needs to be reconsidered in future research: The missing correlation E ~ SMS shows that the rationale that led to their differentiation seems to hold (see 3.2). The most prominent difference between elementary *fluency* and teenage *proficiency* concerns T: once a threshold of, say, T = 4 to 4.5 is passed, it is irrelevant. With respect to AGE and GRADE we tend to conclude that oral reading proficiency is – probably to a much greater extent as, say, mathematical proficiency – more a matter of individual talent than of age or learning advancement. The second question about relations between perceptual parameters and parameters that can be measured in the signal can only now that a perceptual foundation is established be envisaged. As elaborated in [12] we see considerable involvement of two measures of intonational style (*wiggliness* and *spaciousness* of the pitch contour) with C. Maybe in future research we can succeed with bringing perception and signal based values even better in line, rendering costly rating procedures unnecessary.

For educational research we can state that more openness to interdisciplinary perspectives would definitely be promising.

6. REFERENCES

[1] 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. Washington, D.C: Government Printing Office.

[2] Fant, G., Kruckenberg, A. 1989. Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR*, *30(2)*, 001-080.

[3] Blaauw, E. 1994. The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. *Speech Communication, 14(4)*, 359–375.

[4] Cowie, R., Douglas-Cowie, E., Wichmann, A. 2002. Prosodic characteristics of skilled reading: Fluency and expressiveness in 8-10-year-old readers. *Language and Speech, 45(1)*, 47–82.

[5] De Ruiter, L. E. 2015. Information status marking in spontaneous vs. read speech in story-telling tasks – Evidence from intonation analysis using GToBI, *Journal of Phonetics*, *48*, 29–44.

[6] Wolters, A.P., Young-Skuk G.K., Szura, J.W. 2020. Is Reading Prosody Related to Reading Comprehension? A Meta-analysis. *Scientific Studies of Reading, 26(1)*, 1-20.

[7] Godde, E., Bosse, M.L., Bailly, G. 2020. A review of reading prosody acquisition and development. *Reading and Writing, 33(2)*, 399-426.

[8] Miller, J., Schwanenflugel, P. J. (2006). Prosody of syntactically complex sentences in the oral reading of young children. *Journal of Educational Psychology, 98(4)*, 839–853.

[9] 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(2)*, 230–251.

[10] Sappok, C, Linnemann, M., Stephany, S. 2020: Leseflüssigkeit – Prosodie – Leseverstehen. Eine Longitudinalstudie zur Entwicklung der Leseflüssigkeit von Jahrgangstufe 3 bis 7. In: Rautenberg, I. (ed): *Evidenzbasierte Forschung zum Schriftspracherwerb*. Baltmannsweiler: Schneider Verlag Hohengehren, 175-209.

[11] Sappok, C. 2021: Exploring Advanced Prosody – eine Best-Practice-Untersuchung zum lauten Lesen in der weiterführenden Schule. In: Gailberger, S., Sappok, C. (eds.): Weiterführende Grundlagenforschung in der empirischen Leseforschung und Lesedidaktik. Theorie – Empirie – Didaktik. SLLD-B 1, 69-81. https://omp.ub.rub.de/index.php/SLLD/catalog/view/189/167/1101.

[12] Wehrle, S., Sappok, C. 2023. Evaluating prosodic aspects of oral reading proficiency in schoolchildren: effects of gender, genre and grade. *Proc. 20th ICPhS*, Prague.

[13] 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

[14] Hallgren, K. A. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol, 8(1)*, 23-34.

[15] Boersma, P., Weenink, D. (2020). Praat: doing phonetics by computer [Computerprogram]. Version 6.1.16. http://www.praat.org/.

[16] Sappok, C., Arnold, D. 2012. More on the Normalization of Syllable Prominence Ratings. In: *Proc. 13th Interspeech*, Portland, Oregon, USA.