INTERNATIONAL LITERACY ASSOCIATION READING RESEARCH QUARTERLY

Growth in Oral Reading Fluency in a Semitransparent Orthography: Concurrent and Predictive Relations With Reading Proficiency in Norwegian, Grades 2–5

Anne Arnesen
Johan Braeken
University of Oslo, Norway

Scott Baker
Southern Methodist University, Dallas, Texas, USA

Wilhelm Meek-Hansen
Terje Ogden
Norwegian Center for Child Behavioral Development, Oslo, Norway

Monica Melby-Lervåg
University of Oslo, Norway

ABSTRACT

This study investigated an adaptation of the Oral Reading Fluency (ORF) measure of the Dynamic Indicators of Basic Early Literacy Skills into a European context for the Norwegian language, which has a more transparent orthography than English. Second-order latent growth curve modeling was used to examine the longitudinal measurement invariance of the ORF measure, the growth in oral reading fluency within and across grades 2-5, the relative stability of the ORF measure, and the relationship between the ORF measure and high-stakes national tests of reading proficiency. Results showed that the ORF passages measured the same underlying construct, but some passages stood out regarding the invariance pattern. The oral reading fluency growth curve models demonstrated a linear growth in grades 2 and 3 and a nonlinear growth in grades 4 and 5. Initial individual differences varied more than growth rates, which for all were positive but largest in grades 3 and 4. High relative stability in the ORF measure was found across grades. The concurrent and predictive relations of the ORF measure on the Norwegian national reading tests were moderate to strong (range = .44-.75). Findings indicated that the ORF is a reliable and valid measure of reading in Norwegian grades 2-5 and easy and fast to administer. The ORF measure might contribute to early identification of students at risk for reading difficulties in an orthography more transparent than English. Implications for school practice and future research are discussed.

In the United States, elementary schools commonly use a measure of reading fluency, called oral reading fluency (ORF), to screen students for reading difficulties and examine their reading progress over time (S.K. Baker et al., 2008; Deno, Mirkin, & Chiang, 1982; Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993; Shinn, 1998). In this approach, students read aloud grade-specific stories in a one-on-one testing setting, and the number of words read correctly in one minute constitutes their reading performance score (Deno et al., 1982; Shinn, 1989, 1998). In a systems-level approach to screening students for reading problems and monitoring their progress over time, the ORF measure is typically administered three times per school year to all students (S.K. Baker et al., 2011; Shinn, 1989; Shinn, Shinn, Hamilton, & Clarke, 2002).

In Europe, however, few systematic studies have been conducted concerning the instruments that schools use to assess reading skills

Reading Research Quarterly, 0(0)
pp. 1–25 | doi:10.1002/rrq.159

and progress. For reading proficiency, there seems to be a large variation in the types of screening instruments that schools use (Standards & Testing Agency, 2016; Statens Beredning för Medicinsk Utvärdering [SBU], 2014). In the United Kingdom, reading assessments have traditionally focused on reading accuracy tests, such as the phonics screening check applied in first grade (see Standards & Testing Agency, 2016). The emphasis on accuracy is likely to be due to the fact that English has an opaque orthography with inconsistent relations between letters and sounds as compared with other European languages. In more transparent European languages, however, assessments of decoding skills have generally focused on reading fluency rather than accuracy (for an overview, see SBU, 2014). To our knowledge, none of the tests used in European settings include monitoring students' progress in reading fluency over time, as is done with the ORF measure. That is, most of the reading fluency measures in European settings are administered as one-shot assessments. This is a serious omission in light of the importance of reading fluency in the overall development of reading proficiency (Kuhn & Stahl, 2003; LaBerge & Samuels, 1974). Also, concerns have been raised about the lack of psychometric validation of the screening tests and their ability to identify struggling readers (Duff, Mengoni, Bailey, & Snowling, 2015; SBU, 2014).

Identifying struggling readers at an early age is important to provide appropriate interventions for these students. Many students fail in developing well-functioning reading skills. For instance, the PISA studies have shown that 24% of the 15-year-old students in the Organisation for Economic Co-operation and Development (OECD) member countries have low performance in reading comprehension (OECD, 2013). This problem is worrisome because reading comprehension is consistently, across many different contexts (e.g., across languages, in many different countries), a strong predictor of learning overall and specific academic outcomes in multiple subjects (García-Madruga, Vila, Gómez-Veiga, Duque, & Elosúa, 2014; Melby-Lervåg & Lervåg, 2014b). Furthermore, because success in education is strongly related to future possibilities and accomplishments for students, promoting students' reading skills is crucial (Gustafsson et al., 2010). Thus, it is prudent to establish practices and systems for screening students for reading problems. This can support data-based decisions for early intervention, and progress monitoring of students' reading proficiency over time can determine whether interventions are having their intended impact.

The purpose of the present study is to examine the psychometric properties of the ORF measure and its relationship with high-stakes reading tests in a large sample of Norwegian students. The ORF measure used is based on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), a measure widely used in the United States. In this study, we adapted the measurement approach for use in Norwegian, a more transparent orthography than English. Only a small number of ORF studies have been conducted in languages other than English. A Spanish ORF measure, also adapted from DIBELS, has been studied in a U.S. educational context on Spanish-speaking immigrant students (D.L. Baker, Stoolmiller, Good, & Baker, 2011). Thus, these results are not very transferable to a European setting with mainly monolingual students. Although a variety of reading fluency measures are used in European countries (see, e.g., Parrila, Aunola, Leskinen, Nurmi, & Kirby, 2005; Veenendaal, Groen, & Verhoeven, 2015), no studies have been conducted using an adaption of the ORF measure based on DIBELS.

Reading Comprehension: The Ultimate Goal of Reading Proficiency

The ultimate purpose of reading is to extract meaning from text—in other words, to read with comprehension. Several theories have been suggested to explain the development of reading and reading comprehension (Cromley & Azevedo, 2007; Hoover & Gough, 1990; Kintsch, 1988; LaBerge & Samuels, 1974; Perfetti & Stafura, 2014). However, in elementary school students, the theoretical foundation known as the simple view of reading has the strongest empirical support (Gough & Tunmer, 1986; Hoover & Gough, 1990). According to this theory, reading comprehension is the product of the ability to decode words and sentences fluently, accurately, and with automaticity and being able to understand the meaning of these words in the context in which they are used. A number of studies have shown that decoding and listening comprehension can explain much of the variation in students' reading comprehension (for a review, see García & Cain, 2014). In fact, in a recent study using latent variables controlling for measurement error, the features of the simple view of reading explained as much as 94% of the variation among students, leaving little variation left to be explained by other variables (Foorman, Koon, Petscher, Mitchell, & Truckenmiller, 2015). Thus, learning to decode accurately and fluently, together with understanding the meaning of words, is paramount for developing well-functioning reading skills.

More specifically, *decoding skills* refers to the ability to accurately and automatically decipher the relationship between letters and sounds in words and sentences. *Reading fluency* is commonly defined as reading with accuracy, speed, and expression or prosody

(Rasinski, Reutzel, Chard, & Linan-Thompson, 2011; Schwanenflugel, Hamilton, Kuhn, Wisenbaker, & Stahl, 2004; Veenendaal et al., 2015). Recognizing and identifying words implies accurate decoding, but decoding is not necessarily dependent on knowing the meaning of the words, because it is possible to decode nonsense words or to decode real words but not understand the words' meanings. However, several foundational theories of decoding suggest that when a student knows the meanings of the words in a text and can activate this from his or her lexicon, words are more likely to be read automatically and fluently (Perfetti, 1985; Seidenberg & McClelland, 1989). This has also solid empirical support; it is easier to read fluently when you know the meanings of the words (see van IJzendoorn & Bus, 1994). Thus, the more automatic decoding skills are, the less attention needs to be used to assist in the decoding process. More resources will then be available to focus on comprehension.

In the development of decoding skills, students first learn to master decoding accuracy at the word level, then transfer these skills to passages and texts, and increasingly build reading fluency with connected text. As students get older, they learn to master accurate and fluent decoding skills both at the word and sentence levels (Landerl & Wimmer, 2008). When this is mastered as students get older, the effect of decoding on reading comprehension decreases, and language comprehension skills account for more of the variance in reading comprehension (García & Cain, 2014; Lervåg & Aukrust, 2010). Notably, cross-language studies have found differences in reading development between orthographies with different degrees of transparency (Caravolas et al., 2012; Caravolas, Lervåg, Defior, Seidlová Málková, & Hulme, 2013). Although the predictors of decoding are similar (Caravolas et al., 2012), the developmental pattern is different, and students learn to decode fluently more slowly in English, as compared with more transparent languages such as Spanish, Czech (Caravolas et al., 2013), and Finnish (Parrila et al., 2005).

ORF as a Measure of Reading Proficiency

An important question concerning the ORF measure has been its association with other measures of reading. There is strong theoretical support for reading fluency as a crucial component in reading comprehension. Pikulski and Chard (2005) described reading fluency as the bridge between decoding and reading comprehension. As mentioned previously, in the United States, the ORF measure is widely used to measure students' growth trajectories in decoding accuracy and automaticity with age-appropriate passages of connected text read aloud. A number of studies (e.g., S.K. Baker et al., 2008; Pikulski & Chard, 2005; Stoolmiller, Biancarosa, & Fien, 2013; Wise et al., 2010) have demonstrated strong correlations between reading fluency and reading comprehension (.60–.90).

Shinn et al. (2002) studied the association between the ORF measure and measures of decoding and of reading comprehension using confirmatory factor analysis. Third- and fifth-grade students were tested on reading tasks, including decoding phonetically regular words and pseudowords, answering literal and comprehension questions, completing cloze items, producing written retells of texts read, and ORF. For the third-grade sample, all measures made a significant contribution to a unitary, reading proficiency model. ORF measures correlated higher with the model than any of the other measures. For the fifth-grade sample, reading proficiency was best characterized as composed of two factors—decoding and comprehension—although these factors were very highly correlated (r = .83). The ORF measure fitted best with the decoding factor but also correlated higher with the comprehension factor than did the literal and inferential comprehension subtests of the Stanford Diagnostic Reading Test. Thus, the ORF measure provides a good index of reading proficiency, including comprehension (S.K. Baker et al., 2008).

The common conceptualization of the positive association between reading fluency and comprehension is that stronger fluency helps free up cognitive resources, which students can then direct toward constructing the meaning of the text. D.L. Baker and colleagues (2011) used structural equation modeling (SEM) to study whether reading with comprehension also has a positive effect on reading fluency. They also asked whether this possible influence might vary depending on the transparency of the language. To study this, reading data were collected in Spanish and English with second-grade English learners being taught to read in both languages. Results showed that ORF had an effect on reading comprehension, but reading comprehension also had an effect on reading fluency. In other words, the association was reciprocal. In addition, the pattern of the associations was the same in English and Spanish. The instructional implications suggest that reading comprehension instruction—teaching students to comprehend text—leads not only to comprehension benefits but also to reading fluency benefits.

Notably, there are also results showing that ORF is a better predictor of reading comprehension than decoding nonsense words (i.e., word attack), decoding real words in word lists, speed of word-reading measures (García & Cain, 2014; Wise et al., 2010), letter naming, vocabulary, or phoneme awareness is

(Kim, Petscher, Schatschneider, & Foorman, 2010). Furthermore, several studies have shown that there is a different set of predictors for decoding word lists versus decoding words accurately and fluently in connected text. The most plausible reason for this is that accurate and fluent text reading is more related to language comprehension, whereas reading decontextualized word lists rests primarily on phoneme awareness, rapid automatized naming, and letter knowledge (D.L. Baker et al., 2011; Hulme, Bowyer-Crane, Carroll, Duff, & Snowling, 2012; Stanovich, 2000). Therefore, when trying to account for students' reading proficiency when they are reading decontextualized word lists versus connected text, it is necessary to consider the reading task, distinguishing between reading word lists and reading words in connected texts (Veenendaal et al., 2015).

ORF as a Measure of Reading Growth Across Time

Another issue in ORF research has been the degree of reading fluency growth over time (Fuchs et al., 1993; Hasbrouck & Tindal, 1992, 2006) and the meaning of that growth in terms of improvements in overall reading proficiency (S.K. Baker et al., 2008). Hasbrouck and Tindal analyzed ORF data collected in the fall, winter, and spring of grades 2–5. Student performance increased over the course of the year as expected, and the cross-sectional data showed that students' reading fluency grew fastest in grades 2 and 3.

Although the majority of ORF studies have been concurrent or cross-sectional, some longitudinal studies have examined predictive relationships over time and estimated the increase in the numbers of words read per week. For instance, Fuchs et al. (1993) conducted the first longitudinal study on ORF. Different students were assessed in grades 1–6, but in each grade, the same students were tested repeatedly over time. Slope of performance was positive in each grade but decreased steadily across grades, consistent with findings reported by Deno et al. (1982) and Hasbrouck and Tindal (1992, 2006). This nonlinear pattern of rapid early growth and later slower growth has been replicated in other studies (S.K. Baker et al., 2008; Nese et al., 2013; Stage & Jacobsen, 2001). Speece and Ritchey (2005) showed that students with high rates of growth on ORF in grade 1 were more likely to maintain strong growth rates in grade 2 and read at grade level at the end of grade 2 than students who had low rates of growth. Using growth curve analysis, Speece and Ritchey also showed that students who were at risk for reading problems at the beginning of first grade had predicted ORF scores at the end of the year that were less than half the magnitude of their peers who were not at risk.

Several longitudinal studies in the United States have shown that growth in ORF is related to reading comprehension within and across school years and grades. For instance, S.K. Baker and colleagues (2008) investigated what unique contribution, if any, slope on ORF made to performance on comprehensive measures of reading. They investigated this with students in grades 1–3 who were tracked longitudinally for either 1.5 years (the middle of first grade to the end of second grade) or for two years (the beginning of second grade to the end of third grade). In each group, ORF data were collected five (first- and second-grade group) or six times (second- and third-grade group), in addition to a pretest and posttest on a comprehensive measure of reading (the SAT–10 or the state reading test). After controlling for initial status on the ORF measure and the comprehensive measure of reading at pretest, slope of ORF still added to the accuracy of predicting performance on the comprehensive measure of reading at posttest. Thus, progress in ORF was positively associated with improvement in reading proficiency. In grades 1–3, Wanzek, et al. (2010) found that ORF was a reliable predictor of student success on two high-stakes national and state-normed measures. Thus, several U.S. studies have provided strong support for the predictive validity of ORF for reading comprehension.

Because ORF is an important developmental indicator of reading proficiency and creates a foundation for reading comprehension (for a review, see Breznitz, 2006), monitoring reading fluency can help schools identify students at risk for reading failure (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Pfost, Dörfler, & Artelt, 2012). By understanding how reading fluency develops and how it in turn relates to reading comprehension, schools can give struggling readers targeted support in the early stages (S.K. Baker et al., 2008; Hosp & Suchey, 2014; Pikulski & Chard, 2005). When examining developmental processes in reading, latent growth curve models offer a particularly useful way to predict and explain change over time (Little, 2013; Rogosa, Brandt, & Zimowski, 1982; Stoolmiller, 1995). Most studies of growth in ORF have used first-order growth models with one indicator of ORF per timepoint. Our approach is to use multiple ORF indicators, which allows for not only a more thorough investigation of the measurement properties of the ORF reading passages in terms of longitudinal invariance but also the use of second-order latent growth models to better account for measurement error in the individual ORF scores (see, e.g., Widaman, Ferrer, & Conger, 2010). At the same time, a second-order latent growth model still also allows for an investigation of the interindividual differences in ORF starting levels, the interindividual differences in ORF growth across the school year, and the relation between these two individual difference factors among students.

Aim of the Study and Research Questions

The overall aim of our study was to examine initial status and growth in ORF and to investigate how ORF relates to students' reading performance on high-stakes national tests focusing on general reading proficiency (decoding and reading comprehension). Using a longitudinal design, we assessed students in grades 2–5 during one school year in Norwegian, a semitransparent orthography. The ORF measure was constructed by developing three unique grade-specific narrative and expository passages to be administered on three measurement occasions per school year (fall, winter, and spring). All passages were constructed to be parallel ORF items, similar in difficulty, but the actual content, in terms of the stories and information presented, differed to avoid practice effects. Assessing the longitudinal measurement invariance of the ORF passages allowed us to determine whether this objective was achieved.

Our study adds to the literature in several ways: As we have seen, the ORF measure is frequently used in the United States, and many studies have shown that it is a valid and reliable index of students' reading development (see, e.g., S.K. Baker et al., 2008; Deno et al., 1982; Fuchs et al., 1993; Good & Kaminski, 2002; Shinn, 1998; Stoolmiller et al., 2013). However, the ORF measure has never previously been adapted to a European setting where there is a variety of reading measures. With exceptions concerning bilingual Spanish-speaking students in the United States, an ORF measure based on DIBELS has not been used in transparent orthographies. Also, in Europe, the lack of psychometric validation of screening measures is a concern (SBU, 2014). Although it is crucial to examine at-risk students' progress from interventions over time, progress monitoring is not integrated in other European reading assessments. Finally, longitudinal invariance is taken for granted. However, it is difficult to design grade-level reading passages that are of comparable difficulty. If ignored, trends in the ORF measure across time might simply reflect a specific performance difference on a specific reading passage, instead of real progress and development.

To add to the previous literature, we will more specifically examine these four research questions:

1. Does ORF measure the same construct over time (i.e., demonstrate measurement invariance)?
2. How much growth do students experience on ORF over the course of the school year?
3. How stable is the rank order among students on ORF over time?
4. What is the association between the ORF measure and high-stakes tests of reading proficiency?

Method

Participants

A total of 2,228 students (48% female) participated in the study. The students were distributed across grades 2–5 in 21 schools across Norway in one school year (2012–2013). The schools were strategically selected to be representative of the Norwegian population. Therefore, they were located in both urban and rural districts across the country, and students from a variety of socioeconomic backgrounds were included. The number of students enrolled in each grade level ranged from four to 73 per school. Each grade level included 557 students on average, and 84% of them were monolingual. Furthermore, 11% of the students had two parents who were both bilingual, and 5% of the students had one parent who was bilingual.

Measures

An overview of the longitudinal study design and timing of the collected measures for each of the four grade levels (2–5) is given in Table 1. This also clarifies the range of predictive and concurrent relations between the ORF measure and the national tests in reading that are possible

TABLE 1
Overview of the Longitudinal Study Design and Timing of the Collected Measures for Grades 2–5

Year	2012–2013			2013–2014
Period	Fall	Winter	Spring	Fall
Grade 2		*Grade 2*		
ORF	1–3	4–6	7–9	
NTRP			Assessment	
Grade 3		*Grade 3*		
ORF	10–12	13–15	16–18	
NTRP			Assessment	
Grade 4		*Grade 4*		*Grade 5*
ORF	19–21	22–24	25–27	
NTRP				Test
Grade 5		*Grade 5*		
ORF	28–30	31–33	34–36	
NTRP	Test			

Note. NTRP = national tests in reading proficiency; ORF = oral reading fluency reading passages.

to assess in this study. Each school's assessment team had a data coordinator who was responsible for entering the data in an Excel spreadsheet established for this study.

ORF

The ORF measure and procedures are based on those of the ORF subtest drawn from the reading assessments, DIBELS sixth edition (Good & Kaminski, 2002). ORF was measured by three grade-specific narrative and expository passages on three measurement occasions at four-month intervals (fall, winter, and spring) during the 2012–2013 school year (see Appendix A for an overview). The range of words in each passage by grade varied: grade 2 = 190–207; grade 3 = 251–299; grade 4 = 297–310; and grade 5 = 300–326. Each passage was read aloud for one minute following standardized procedures. A trained teacher administered the ORF measure in an individual setting. Students were asked to read the passages aloud as accurately and as best they could until the teacher told them to stop. Students were told that if they got stuck, the teacher would tell them the word so they could keep reading. Words self-corrected within three seconds were scored as accurate. For each of the three passages, the number of words read correctly in one minute was the ORF raw score used in data analysis.

The ORF measures were administered individually to students by a teacher who was part of an assessment team that was established in each school for the purpose of the study. The assessment team consisted of expert teachers in reading, classroom teachers, or special teachers employed in the schools. All teachers who administered the ORF assessments received half-day training in the procedures of administration and scoring. For each grade level in this study, three new reading passages were administered at each measurement occasion. The full set of 36 reading passages (four grade levels × three passages × three occasions) were specifically developed in Norwegian for grades 2–5. Each set of the nine grade-level passages was constructed so each passage was similar to the others in the set in terms of purpose and passage characteristics, such as difficulty, length, and format. According to standard administration of ORF passages (Good & Kaminski, 2002), students who read fewer than 10 words correctly on the first of the three passages were not administered passages 2 and 3. In such cases, the ORF raw score for the latter two passages is not recorded and is therefore missing by design.

In this study, the alternate-form reliabilities were very high for all of the ORF passages within and across grades 2–5, ranging from .92 to .97 (see Table 2). This is in line with reliability findings in U.S.-based studies, where similar reliabilities have been reported as ranging from .89 to .97 for ORF measures (see, e.g., Cummings, Biancarosa, Schaper, & Reed, 2014; Good, Kaminski, & Dill, 2002; Stoolmiller et al., 2013).

National Tests of Reading Proficiency (NTRP)

In Norway, there are two types of national tests of reading proficiency administered to students in elementary school. The first type targets the early grades (1–3) and is a mandatory reading assessment for use in all Norwegian schools. It is group administered annually in the spring and aims to identify the need for support at both the individual and school levels. The second type is used in grade 5 only and functions as part of the quality assessment system for the Norwegian schools. This national test is group administered annually in the fall to all Norwegian students in grade 5. Each year, a new version of the NTRP is developed for both test types (Norwegian Reading Centre, 2013a, 2013b; Skaftun, Stangeland, Solheim, & Mangen, 2013; Solheim, Skaftun, & Walgermo, 2012). For this study, the annual updating of the NTRP implies that the measures differ across the four grade levels in terms of complexity.

For grade 2, the national reading assessment in spring 2013 consisted of the following seven subtests (see Appendix B for descriptions of the subtests): recognizing letters, writing words, reading words, splitting compound words, reading sentences, following written instructions, and reading text. For grade 3, the national reading assessment in spring 2013 consisted of the following four subtests (see Appendix B for descriptions): chains of words, reading narrative text, word knowledge, and reading expository text. Because no NTRP is available for grade 4, the national reading test score is based on the fall 2013 version of the test from when the fourth-grade students moved to grade 5; for grade 5, the score is based on the fall 2012 version (Cronbach αs based on official population data for the two fifth-grade tests are 0.86 and 0.86, respectively). The fifth-grade tests consisted of multiple texts to assess students' decoding and comprehension skills. Test formats included multiple-choice, closed-ended, and open-ended questions. Students had to find information in the texts, interpret the texts, and explain the meaning of them. The test used in 2012 consisted of 28 items, and the test used in 2013 consisted of 29 items.

Data Analysis

For each grade level (2–5), statistical models for data analysis were established in line with the longitudinal study design and within a SEM framework using the lavaan package (Rosseel, 2012) in the statistical software environment R. Full information maximum likelihood was used to handle missing data and make use of all available information for each individual. We applied robust (Huber–White) standard errors for all estimated parameters and a scaled goodness-of-fit chi-square for statistical inference. Model fit was evaluated based on commonly recommended goodness-of-fit indexes (Hu & Bentler,

TABLE 2
Descriptive Statistics for All Reading Passages (RP) That Form the Basis of the Oral Reading Fluency Measure Across Grades 2–5

Oral reading fluency: Grade 2									
	Fall (reliability = .97)			**Winter (reliability = .96)**			**Spring (reliability = .96)**		
	RP 1	**RP 2**	**RP 3**	**RP 4**	**RP 5**	**RP 6**	**RP 7**	**RP 8**	**RP 9**
M	37.78	44.21	37.63	53.81	52.12	53.43	69.18	63.88	62.21
SD	27.96	27.12	24.37	32.35	29.32	29.23	31.39	32.20	31.50
[min, max]	[0, 162]	[2, 162]	[3, 141]	[4, 167]	[3, 162]	[8, 166]	[4, 190]	[3, 187]	[3, 201]
Skewness	1.00	1.15	1.13	0.74	0.91	0.79	0.63	0.59	0.77
Kurtosis	3.91	4.28	4.22	2.87	3.70	3.37	3.17	3.05	3.58
n	411	373	372	466	462	461	459	459	458
Oral reading fluency: Grade 3									
	Fall (reliability = .96)			**Winter (reliability = .95)**			**Spring (reliability = .94)**		
	RP 10	**RP 11**	**RP 12**	**RP 13**	**RP 14**	**RP 15**	**RP 16**	**RP 17**	**RP 18**
M	66.72	73.82	69.64	84.19	82.66	82.54	96.20	88.89	95.46
SD	33.78	37.53	36.10	33.16	33.82	33.24	35.10	33.99	34.98
[min, max]	[0, 184]	[0, 198]	[0, 196]	[9, 190]	[10, 198]	[13, 220]	[13, 220]	[10, 229]	[11, 230]
Skewness	0.65	0.52	0.63	0.26	0.15	0.41	0.47	0.54	0.24
Kurtosis	3.25	2.86	3.00	2.94	2.68	2.93	3.25	3.63	3.42
n	435	434	432	472	472	472	471	471	471
Oral reading fluency: Grade 4									
	Fall (reliability = .93)			**Winter (reliability = .93)**			**Spring (reliability = .95)**		
	RP 19	**RP 20**	**RP 21**	**RP 22**	**RP 23**	**RP 24**	**RP 25**	**RP 26**	**RP 27**
M	98.70	109.61	95.59	100.10	122.14	102.86	123.62	128.59	112.90
SD	34.26	37.32	36.59	35.11	38.29	32.38	37.61	35.71	37.88
[min, max]	[4, 194]	[6, 214]	[5, 203]	[16, 204]	[17, 213]	[20, 199]	[22, 221]	[28, 232]	[25, 208]
Skewness	−0.20	−0.11	−0.12	0.21	−0.14	0.20	−0.08	−0.13	0.16
Kurtosis	2.77	2.76	2.84	2.69	2.74	2.90	2.82	3.23	2.57
n	475	475	475	532	533	532	443	443	441
Oral reading fluency: Grade 5									
	Fall (reliability = .92)			**Winter (reliability = .93)**			**Spring (reliability = .94)**		
	RP 28	**RP 29**	**RP 30**	**RP 31**	**RP 32**	**RP 33**	**RP 34**	**RP 35**	**RP 36**
M	107.95	103.52	120.00	117.88	126.11	128.49	122.59	118.37	117.79
SD	29.30	33.36	37.04	31.66	34.78	30.79	31.25	32.97	36.67
[min, max]	[19, 193]	[20, 184]	[25, 213]	[30, 205]	[24, 214]	[32, 224]	[26, 27]	[28, 227]	[15, 226]
Skewness	0.06	−0.10	−0.10	−0.09	−0.17	−0.25	−0.06	0.02	−0.03
Kurtosis	3.06	2.41	2.53	2.81	2.68	3.00	3.17	3.05	2.61
n	461	461	461	482	482	482	443	443	443

Note. M = mean; *SD* = standard deviation. Reliability was measured by calculating the mean of the correlations between the passages at the timepoint and the following timepoints.

1999), including the chi-square test of exact model fit, the root mean square error of approximation (RMSEA: ≤0.08 = acceptable, ≤0.05 = good) to assess close fit, the comparative fit index (CFI: ≥0.95 = good) contrasting to a null independence model, and the standardized root mean square residual (SRMR: ≤0.05 = good).

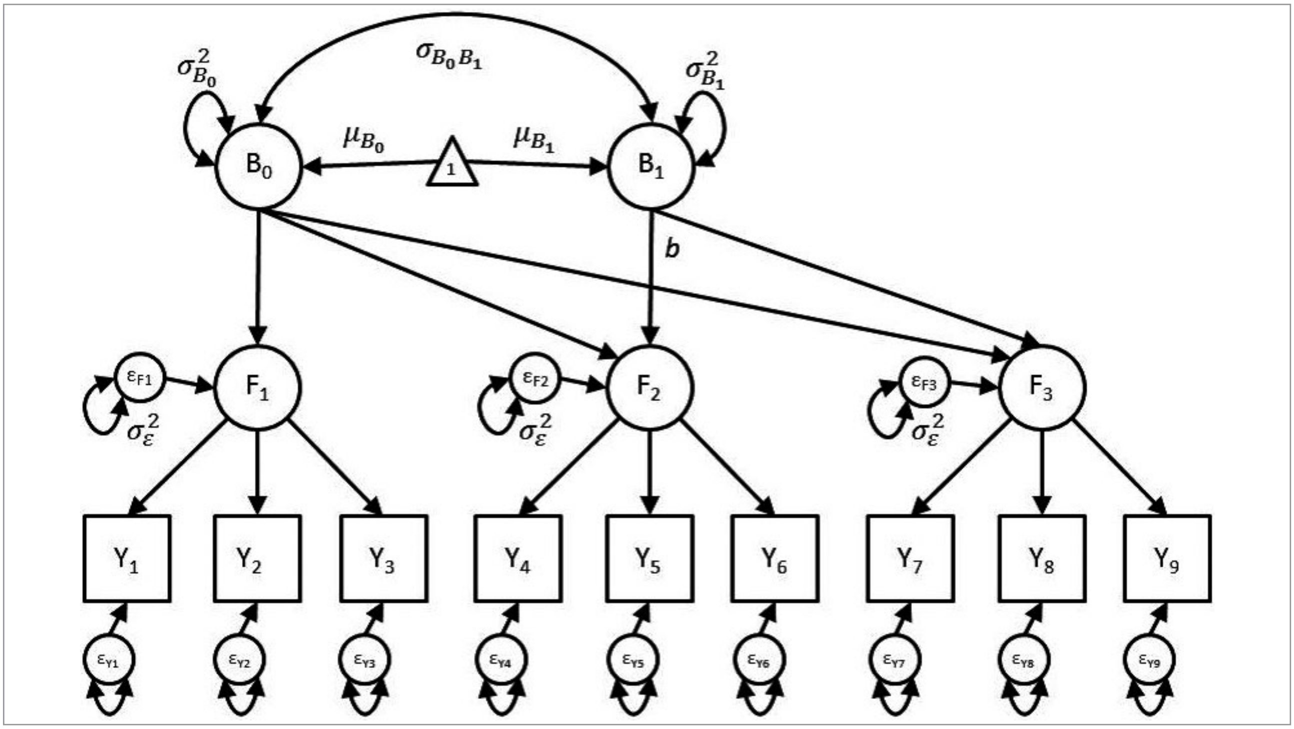
Longitudinal Measurement Invariance

Although the ORF reading passages were designed to be of comparable difficulty, we first verified this design feature for each grade level by investigating the longitudinal measurement invariance of the latent variable measurement model with all nine reading passages. We followed a model comparison procedure (for an overview, see Millsap, 2011) assessing the viability of restricting specific model parameters to be parallel invariant across the nine reading passages. When full invariance was not obtainable, we aimed to establish partial invariance by freeing up some invariance constraints under the condition that at each measurement timepoint, at least one reading passage was kept parallel invariant. The reason to pursue (at least partial) invariance is that otherwise differences in the ORF measure across time might simply reflect an idiosyncratic performance difference on a specific reading passage (e.g., an intended parallel-designed passage that unexpectedly turns out to be extremely difficult or easy in practice might overrule the general ORF trend across time). Model comparison is based on assessment of absolute goodness of fit and on the relative fit profile across the sequence of invariance models as indicated by differences in fit indexes such as the CFI (see, e.g., Cheung & Rensvold, 2002; Little 2013).

ORF Growth Models

Having established longitudinal (partial) invariance, a growth model is posited on top of the latent ORF factors in the measurement model. A path diagram of such a second-order latent growth model is given in Figure 1. The second-order latent growth model not only accounts for the measurement error in the individual ORF reading passage scores but also allows us to investigate the interindividual differences in ORF starting levels of the students in the same grade level (i.e., as indicated by the variance $\sigma^2_{B_0}$ of the random intercept factor B_0), the interindividual differences in ORF growth across the school year (i.e., as indicated by the variance $\sigma^2_{B_1}$ of the random slope factor B_1), and the relation between these two individual difference factors (i.e., as indicated by their covariance $\sigma_{B_0B_1}$). The mean parameter μ_{B_0} of the random intercept factor represents the average ORF starting level. To allow for a potential

FIGURE 1
Path Diagram of the Second-Order Latent Growth Model for the Nine (3 × 3) Parallel-Invariant Oral Reading Fluency Passages in Each of the Grades (2–5)



Note. Observed variables are represented by squares, latent variables by circles, and constants by triangles. The value of paths corresponding to nonannotated directed arrows is fixed at 1.

nonlinear growth trajectory, the loading for the in-between timepoint on the random slope factor B_1 is estimated freely, allowing the corresponding parameter b to be interpreted as the proportional change in ORF relative to the average change in ORF from the first timepoint to the last timepoint as represented by the mean parameter μ_{B_1} of the random slope factor. Variances of the residual time-specific ORF factors are constrained to be equal (i.e., parameter σ_{ε}^2).

Note that partial invariance would imply that some of the observed reading passage scores do not follow the general implied growth curve trend exactly and show a slightly differential pattern in either the observed mean score, as reflected in a nonzero intercept parameter (i.e., an additional direct path from the constant to one of the squares), or in the observed score (co)variance, as reflected in a freely estimated loading of the reading passage on its corresponding time-specific ORF factor.

The ORF Measure and the NTRP

The periodic changes in the Norwegian NTRP prevent a clear-cut comparison across time of the NTRP scores and of their link with the ORF measures. Yet, it results in the added benefit of having a variety of reading proficiency measures to evaluate the value of the ORF measure against it (see Table 1).

Missing Data Analysis

In general, missingness and dropout can be expected in every longitudinal study. Yet, its impact depends on whether data are systematically missing according to processes that can bias the measures of interest (e.g., only low-scoring students dropping out) or whether missingness is merely due to some random idiosyncratic events or planned because of the design. Random events here are relocation of students, students or administering teachers being absent due to illness, and practical administration issues preventing two schools from conducting data collection at the first timepoint for grade 2 and one school not completing data collection at the second and third timepoints in any grade. For grade 4 specifically, national reading examination tests were not available at the time of the ORF measure's administration but only one year after, when students moved to grade 5, such that there was less incentive for local data coordinators to follow through with delivering this extra set of NTRP data for all students.

Initial exploratory analyses indicate that having one or more missing scores at the later two timepoints is not related to performance at the first ORF timepoint. Given the low stakes of the ORF assessment, it is reasonable to assume that missingness is indeed random and not due to expected negative consequences of the ORF assessment for schools, teachers, or students involved. There is one design factor present: 37 of 528 students in grade 2 reading the first of three ORF passages at the first timepoint with less than 10 words read correctly per minute were exempted from the remaining two passages for that timepoint, in line with the ORF measure's administration protocol. Only one or two such cases occurred in later measurement occasions and in later grades.

A complete set of nine ORF scores was available for 308 students in grade 2 (58%; 30% missing between one and three scores, 12% missing more than three scores), 362 students in grade 3 (66%; 22% missing between one and three scores, 12% missing more than three scores), 413 students in grade 4 (70%; 7% missing between one and three scores, 23% missing more than three scores), and 384 students in grade 5 (68%; 11% missing between one and three scores, 21% missing more than three scores). National test scores in reading proficiency were available for 384 students in grade 2 (73%), 351 students in grade 3 (64%), 165 students in grade 4(28%), and 302 students in grade 5 (53%). The missingness in measures was partially overlapping, with 247 (47%), 280 (51%), 110 (19%), and 239 (42%) students having outcomes both on all ORF measures and on all NTRP in grades 2–5, respectively. This implies that when taking into account these practical data collection limitations, a very conservative estimate of the effective sample sizes in the different grades still amounts to about 250, which provides a large enough data coverage base for analysis of ORF–NTRP interrelations (for grade 4, standard errors can be expected to be slightly larger due to the relatively smaller complete overlap). An overview of the sample size for each measure across the year per grade is available in Tables 5–8.

Results

Descriptive Statistics

The ORF Measure

Descriptive statistics for all reading passages that form the basis of the ORF measure across the four grade levels are presented in Table 2. It is readily apparent that mean performance in the number of words read correctly per minute increases gradually within each grade level across the year and also across grades, although this pattern becomes less pronounced when comparing grades 4 and 5. The standard deviations within and across grades are rather similar and large, indicating a similar spread of scores across grades and measurement timepoints and large individual differences across students. Patterns of higher scores as students move up in grade with smaller differences at higher grades, and a relatively consistent spread among students across grades with somewhat larger standard deviations in the upper grades, is consistent with previous research on ORF. Less variability among standard deviations might be interpreted positively because it shows similarity in the spread of scores. This finding is also consistent with previous studies on ORF in English-speaking students (see, e.g., S.K. Baker et al., 2008). Skewness and kurtosis statistics are within acceptable ranges for further SEM.

NTRP

Descriptive statistics for all NTRP across the four grade levels are presented in Table 3. The scores on the national reading test are higher in grade 4 than grade 5, but scores on the two versions of this test are not directly comparable because the content of the subtests changes from year to year. The sample descriptive statistics for the fifth graders on the national reading tests map closely to official population statistics, a finding that further supports the representativeness of the study sample. For the national reading assessments in the lower grades, no official statistics were available. For the NTRP scores, skewness and kurtosis statistics are also within acceptable ranges for further SEM, except for the first subtest in grade 2. Due to the clear ceiling effect on this "recognizing letters" measure (i.e., almost all students obtain the maximum score of 25), this subtest will not be considered in further analyses.

Longitudinal Measurement Invariance

Investigating change in ORF across time and interrelations across time with external variables such as those of the NTRP requires that we have measured the same ORF construct with the same metric at each occasion. Because three ORF scores are available at each occasion, we can explicitly evaluate this required longitudinal measurement invariance. If the ORF measurement instrument does not exhibit evidence of longitudinal invariance, then the interpretation of change in mean scores and correlations between timepoints may be ambiguous (Horn & McArdle, 1992).

Table 4 provides an overview of the measurement invariance model results, treating all reading passages within a grade level as parallel ORF indicators. In each grade (2–5), the configural reference model (Horn & McArdle, 1992; Little, 2013) provided an excellent goodness of fit to the data, reflecting that the nine ORF passages were measuring the same underlying construct. Restricting the loadings of the ORF passage scores to be equal across time had only a small impact on the resulting fit to the data. This implies that the assumption of metric invariance was met such that latent ORF scores can be considered to be expressed in the same units across time. Restricting the intercepts of the ORF scores to be equal across time had a dramatic impact on the resulting fit to the data. This indicates that although all reading passages were designed to be comparable in principle, there were particular passages that stood out empirically and biased the general trend in ORF latent means across time. Yet, by relaxing some of the restrictions for these differentially functioning reading passages, a well-fitting partial scalar invariance model was still obtained for every grade level, allowing for meaningful unambiguous comparisons and further analyses of ORF across time.

Growth in ORF

Having established longitudinal partial invariance in each grade level (2–5), a growth model was posited on top of the latent ORF factors in the measurement model (see Figure 1) of each grade level. The resulting second-order latent growth models showed good fit to the data: Grade 2: $\chi^2(33) = 81.17$, $p < .001$, CFI = 0.991, RMSEA = 0.053, $p = .335$, SRMR = 0.022; grade 3: $\chi^2(36) = 118.53$, $p < .001$, CFI = 0.984, RMSEA = 0.065, $p = .018$, SRMR = 0.041; grade 4: $\chi^2(32) = 157.28$, $p < .001$, CFI = 0.977, RMSEA = 0.082, $p = .082$, SRMR = 0.053; and grade 5: $\chi^2(31) = 314.00$, $p < .001$, CFI = 0.959, RMSEA = 0.094, $p = .128$, SRMR = 0.094.

Average Growth Trajectory

Figure 2 provides an overview of the estimated average ORF growth trajectory across grades 2–5 if we examine the results of the four grades together. The Norwegian students began the year reading an average of about 38, 66, 97, and 104 words correct per minute (WCPM) in grades 2–5, respectively. The average growth in number of WCPM was about 26, 31, 26, and 14 in grades 2–5, respectively. The average peak performance in the growth trajectories was 65, 97, 123, and 129 WCPM in grades 2–5, respectively.

The students in grade 2 started off reading about 38 WCPM (i.e., random intercept mean $\mu_{B_0} = 38.29$ [1.27], $p < .001$), which rapidly increased (i.e., random slope $\mu_{B_1} = 26.11$ [0.70], $p < .001$) across the year up to about 65 WCPM in the spring. The growth trajectory is approximately linear, with 57% (i.e., loading $b = 0.57$ [0.02]) of the total average change in ORF in grade 2 already occurring by winter. A similar pattern of results occurred in grade 3 ($\mu_{B_0} = 66.42$ [1.52], $p < .001$; $\mu_{B_1} = 31.48$ [0.77], $p < .001$; $b = 0.58$ [0.02]). In grade 4, the growth trajectory starts at about the same level ($\mu_{B_0} = 97.39$ [1.44], $p < .001$) as the spring ORF results for grade 3 but still shows continuing ORF growth ($\mu_{B_1} = 25.92$ [0.76], $p < .001$), although initially there is now a slower increase between fall and winter ($b = 0.15$ [0.03]), with an increase to spring accounting for 85% of the average total growth. In grade 5, the ORF growth trajectory seems to decrease ($\mu_{B_1} = 14.13$ [0.84], $p < .001$), with the initial average level in the fall for grade 5 ($\mu_{B_0} = 104.36$ [1.42], $p < .001$) being in the zone of the winter results for grade 4. The growth trajectory in grade 5 is no longer systematically increasing, with the peak ORF performance occurring in the winter ($b = 1.74$ [0.08]) and not in the spring as would be expected.

TABLE 3
Descriptive Statistics for National Tests of Reading Proficiency (NTRP) Across Grades 2–5

	Grade 2								Grade 3			Grade 4	Grade 5
	Recognizing letters	Writing words by listening and spelling	Reading words	Splitting compound words	Reading sentences	Following written instructions	Reading text	Chains of words	Reading narrative text	Word knowledge	Reading expository text	NTRP 2013	NTRP 2012
Timepoint	Spring								Spring			Fall of next year	Fall
M	24.71	13.46	15.43	12.92	14.53	7.86	3.35	24.29	5.33	14.97	3.99	21.22	18.57
SD	1.27	2.26	4.35	5.59	3.74	2.54	1.47	9.04	2.30	3.94	1.69	6.25	6.32
[min, max]	[24, 25]	[5, 21]	[4, 21]	[0, 21]	[2, 18]	[0, 10]	[0, 8]	[3, 65]	[0, 14]	[2, 20]	[0, 7]	[0, 33]	[0, 35]
Skewness	−6.08	−1.06	−0.37	0.01	−0.89	−1.30	−0.40	0.41	−0.18	−1.09	−0.01	−0.47	−0.40
Kurtosis	45.08	4.36	2.17	1.88	2.86	3.98	2.93	3.60	2.74	3.76	2.53	2.88	2.52
n	384	384	384	384	384	384	384	352	352	351	351	165	302
Official population statistics													
M												21.5	18.50
SD												6.70	6.20
Cronbach's α												0.86	0.86
N												55,272	54,296

Note. M = mean; *SD* = standard deviation. Due to the clear ceiling effect in the first test score of grade 2 (i.e., almost everyone obtains the maximum score), this "recognizing letters" measure will not be considered in further analyses.

TABLE 4
Oral Reading Fluency Longitudinal Measurement Invariance Results for Grades 2–5

Measurement invariance model	χ^2	*df*	*p*	Comparative fit index (CFI)	Root mean square error of approximation	*p*	Standardized root mean square residual	ΔCFI
Grade 2								
Configural	33	24	.102	0.998	0.027	.983	0.004	—
Metric	146	30	<.001	0.978	0.086	<.001	0.050	0.020
Scalar	562	36	<.001	0.902	0.167	<.001	0.064	0.096
Partial	64	31	<.001	0.994	0.045	.684	0.012	0.004
Grade 3								
Configural	17	24	.833	1.000	0.000	1	0.003	—
Metric	77	30	<.001	0.991	0.054	.307	0.036	0.009
Scalar	347	36	<.001	0.939	0.127	<.001	0.048	0.061
Partial	112	34	<.001	0.985	0.065	.021	0.037	0.015
Grade 4								
Configural	56	24	<.001	0.994	0.048	.571	0.005	—
Metric	186	30	<.001	0.971	0.094	<.001	0.063	0.023
Scalar	1,268	36	<.001	0.772	0.242	<.001	0.108	0.222
Partial	138	30	<.001	0.980	0.079	<.001	0.038	0.014
Grade 5								
Configural	47	24	.004	0.997	0.041	.785	0.005	—
Metric	328	30	<.001	0.957	0.133	<.001	0.093	0.040
Scalar	1,072	36	<.001	0.850	0.227	<.001	0.111	0.147
Partial	142	29	<.001	0.984	0.084	<.001	0.070	0.013

Note. In line with the intended oral reading fluency test design, the measurement invariance models treat all reading passages (RPs) as parallel items. Freed invariance constraints for the grade 2 partial model: Loading RP 3 and RP4 and intercept RP 2, RP 3, and RP 7; freed invariance constraints for the grade 3 partial model: Intercept RP 11 and RP 17; freed invariance constraints for the grade 4 partial model: Loading RP 23 and RP 24 and Intercept RP 20, RP 23, RP 26, and RP27; freed invariance constraints for the grade 5 partial model: Loading RP 28, RP 32, and RP 36 and intercept RP 30, RP 31, RP 32, and RP 34.

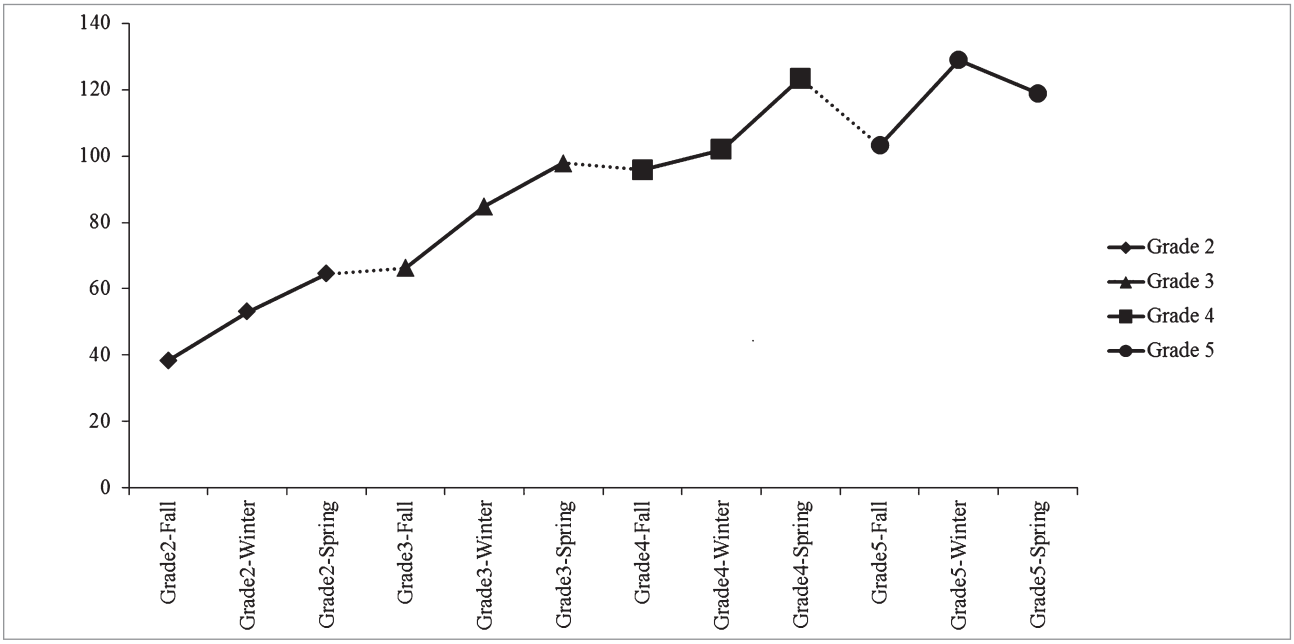
Individual Differences in ORF Development

The boxplots in Figure 3 provide an overview of the individual differences in estimated initial ORF levels and ORF growth rates (i.e., random intercept and slope, B_0 and B_1) in the four grade levels. As expected, initial levels ($\sigma^2_{B_0}$ = 706.39 [59.36], 1,145.34 [75.90], 1,088.59 [64.58], and 967.97 [61.65], respectively) vary much more than growth rates ($\sigma^2_{B_1}$ = 31.86 [31.30], 65.50 [29.61], 59.59 [19.24], and 1.45 [4.77], respectively) across individuals at all grade levels. Estimated population variation in the growth rate across individuals is larger in grades 3 and 4, whereas in grades 2 and 5, the variance could not be estimated very precisely and is smaller (grade 2) to almost nonexistent (grade 5). For grades 3 and 4, there is a small correlation between initial level and growth rate ($r_{B_0B_1} = -.16$, $p = .177$; $r_{B_0B_1} = .289$, $p = .001$). For grades 2 and 5, interpreting a correlation in the presence of a lack of variation of one of its components is not informative. The spaghetti plot in Figure 4 presents the resulting estimated individual growth trajectories. Consistent with students' natural development of reading skills, growth rates are positive for all individuals (sample minimum of estimated growth rates = 18.82, 14.54, 5.34, and 12.02 for grades 2–5, respectively).

Relative Stability of the ORF Measure and Concurrent and Predictive Relations Between the ORF Measure and the NTRP

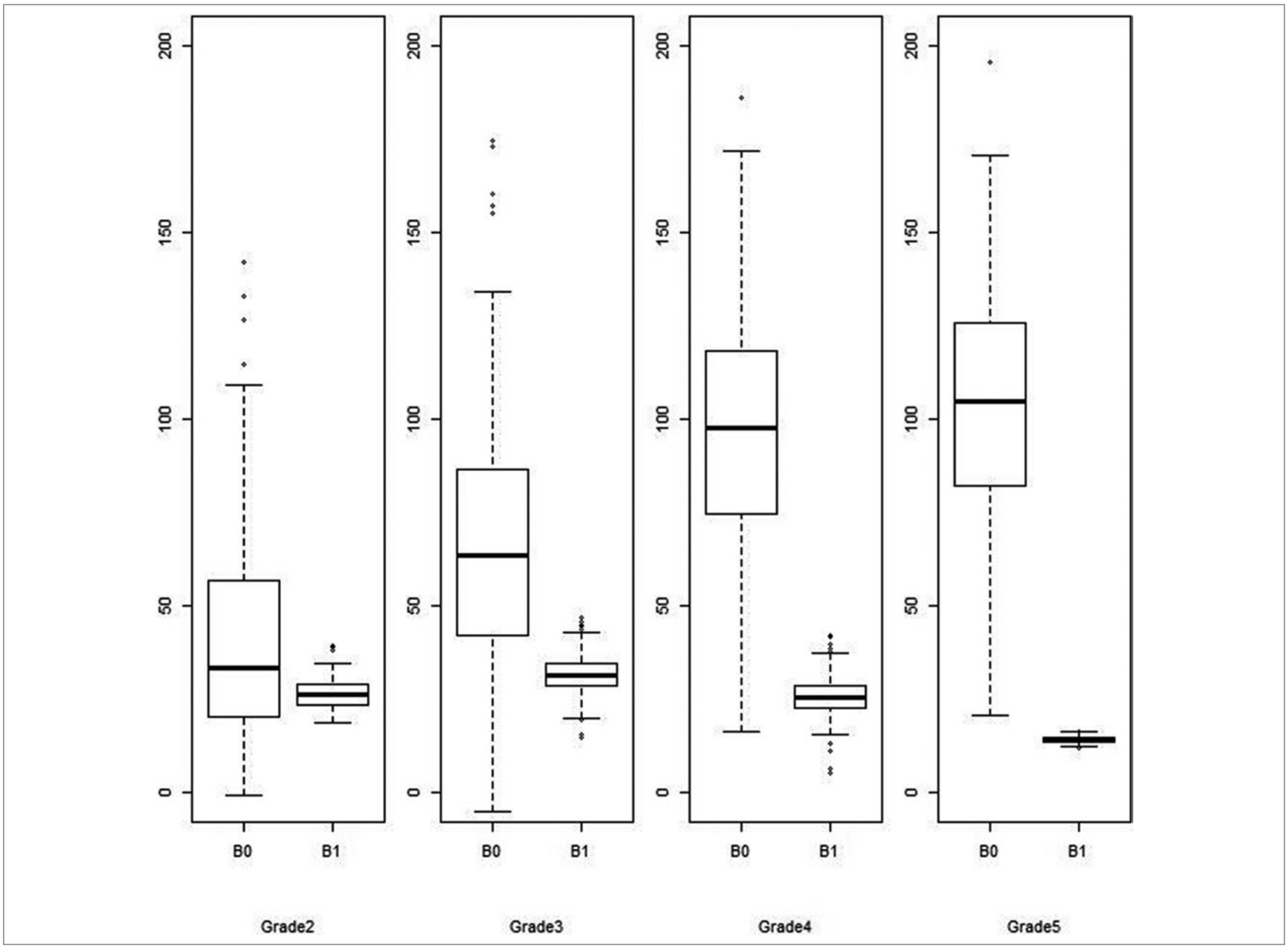
The relative stability of the ORF measure was high, as reflected by correlations of above .9 between the three

FIGURE 2
Oral Reading Fluency Growth Curve Across Grades 2–5



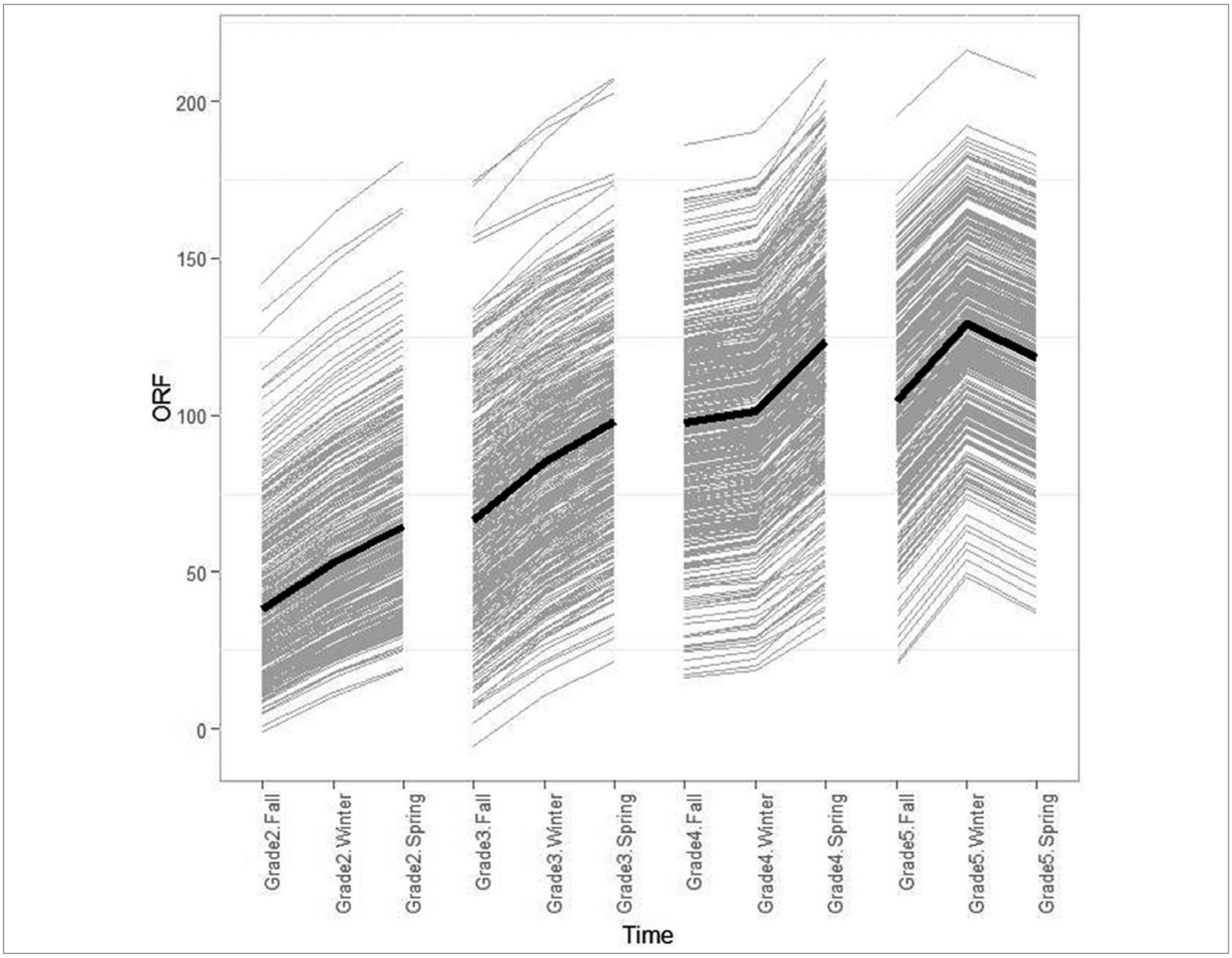
Note. The dotted line indicates the transition from oral reading fluency data based on one grade to another.

FIGURE 3
Individual Differences in Estimated Oral Reading Fluency Growth Parameters Across Grades 2–5



Note. The random intercept factors B_0 represent the starting level, and the random slope factors B_1 represent the growth rate.

FIGURE 4
Estimated Oral Reading Fluency Growth Trajectories Across Grades 2–5



Note. Thin, gray lines represent individuals, and thick, black lines represent grade averages.

ORF factors across measurements within a grade (i.e., r_{F_{t-1},F_t} = .92–.94, .94–.96, .95–.97, and .95–.97, for grades 2–5, respectively). Hence, although ORF increases across the school year in an absolute sense, the relative rank ordering in terms of the students' ORF did not change much (see Figure 4).

The correlations between the ORF measures and the NTRP for grades 2–5 are shown in Tables 5–8. Grade 2 students were administered a national reading assessment consisting of seven subtests in the spring, which allows us to asses both predictive relations with the ORF measure (winter and fall measurement occasions) and concurrent relations with it (spring occasion). The first subtest, recognizing letters, is uninformative because almost all students earn the maximum score, and was consequently dropped from further analyses. The six remaining subtests, which required more elementary operations or targeted subskills needed for reading fluency, more strongly related to the ORF measures (reading words: *r* = .68, .69, and .73, respectively; splitting compound words: *r* = .69, .70, and .75, respectively; reading sentences: *r* = .63, .67, and .73, respectively) than did the subtests requiring more complex operations or higher level skills (writing words by listening and spelling: *r* = .48, .48, and .52, respectively; following written instructions: *r* = .57, .61, and .66, respectively; reading text: *r* = .49, .47, and .53, respectively). Concurrent correlations (i.e., the third *r* value indicative of the spring measurement occasion) were slightly larger than predictive relations (i.e., the first two *r* values indicative of the fall and winter measurement occasion), with a noticeable increase in the relation between the ORF measure and the reading sentences subtest.

Grade 3 students were administered a national reading assessment consisting of four subtests in the spring, which allows us to asses both predictive relations with the ORF measure (winter and fall measurement

TABLE 5
Correlations Between All Observed Measures With the Sample Size at All Timepoints for Grade 2

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. ORF 1	411															
2. ORF 2	**.97**	373														
3. ORF 3	**.97**	**.97**	372													
4. ORF 4	**.89**	**.89**	**.89**	466												
5. ORF 5	**.89**	**.90**	**.89**	**.97**	462											
6. ORF 6	**.89**	**.90**	**.90**	**.96**	**.97**	461										
7. ORF 7	**.89**	**.89**	**.88**	**.90**	**.90**	**.91**	459									
8. ORF 8	**.89**	**.88**	**.88**	**.91**	**.90**	**.91**	**.96**	459								
9. ORF 9	**.89**	**.89**	**.88**	**.91**	**.90**	**.91**	**.96**	**.97**	458							
10. NTRP 1	.10	.08	.07	.14	.13	.14	.18	.19	.17	384						
11. NTRP 2	.48	.42	.38	.48	.47	.47	.53	.52	.52	**.17**	384					
12. NTRP 3	.66	.61	.60	.68	.67	.67	.73	.72	.72	**.29**	**.46**	384				
13. NTRP 4	.67	.63	.63	.71	.69	.70	.74	.73	.75	**.20**	**.46**	**.75**	384			
14. NTRP 5	.62	.55	.56	.66	.65	.67	.72	.72	.71	**.27**	**.52**	**.77**	**.75**	384		
15. NTRP 6	.55	.49	.47	.60	.59	.59	.67	.65	.64	**.21**	**.46**	**.63**	**.60**	**.74**	384	
16. NTRP 7	.48	.38	.40	.47	.45	.46	.54	.53	.50	**.22**	**.46**	**.43**	**.40**	**.49**	**.55**	384

Note. NTRP = national tests of reading proficiency; ORF = oral reading fluency measure. The bold numbers are correlations within construct. Correlations greater than .13 are significant at the .05 level. The diagonal is the sample size at each measure.

TABLE 6
Correlations Between All Observed Measures With the Sample Size at All Timepoints for Grade 3

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13
1. ORF 1	435												
2. ORF 2	**.96**	434											
3. ORF 3	**.96**	**.96**	432										
4. ORF 4	**.91**	**.90**	**.90**	472									
5. ORF 5	**.90**	**.90**	**.90**	**.95**	472								
6. ORF 6	**.90**	**.90**	**.90**	**.95**	**.96**	472							
7. ORF 7	**.89**	**.89**	**.89**	**.91**	**.91**	**.91**	471						
8. ORF 8	**.88**	**.89**	**.89**	**.90**	**.91**	**.91**	**.94**	471					
9. ORF 9	**.88**	**.88**	**.89**	**.91**	**.91**	**.91**	**.95**	**.94**	471				
10. NTRP 1	.72	.71	.68	.72	.75	.72	.74	.70	.73	352			
11. NTRP 2	.43	.45	.45	.52	.50	.52	.48	.48	.50	**.39**	351		
12. NTRP 3	.27	.28	.28	.35	.34	.37	.34	.33	.34	**.29**	**.45**	351	
13. NTRP 4	.36	.39	.40	.45	.45	.46	.46	.45	.45	**.35**	**.49**	**.42**	351

Note. NTRP = national tests of reading proficiency; ORF = oral reading fluency measure. The bold numbers are correlations within construct. All correlations are significant at the .01 level. The diagonal is the sample size for each measure.

TABLE 7
Correlations Between All Observed Measures With the Sample Size at All Timepoints for Grade 4

Measure	1	2	3	4	5	6	7	8	9	10
1. ORF 1	476									
2. ORF 2	**94**	475								
3. ORF 3	**.93**	**.94**	475							
4. ORF 4	**.89**	**.90**	**.91**	532						
5. ORF 5	**.92**	**.92**	**.92**	**.93**	533					
6. ORF 6	**.88**	**.89**	**.90**	**.92**	**.93**	532				
7. ORF 7	**.90**	**.89**	**.89**	**.91**	**.92**	**.90**	443			
8. ORF 8	**.90**	**.89**	**.89**	**.90**	**.91**	**.90**	**.95**	443		
9. ORF 9	**.88**	**.88**	**.89**	**.91**	**.90**	**.91**	**.94**	**.94**	441	
10. NTRP	.52	.48	.50	.50	.48	.46	.48	.47	.50	165

Note. NTRP = national tests of reading proficiency; ORF = oral reading fluency measure. The bold numbers are correlations within construct. All correlations are significant at the .01 level. The diagonal is the sample size for each measure.

TABLE 8
Correlations Between All Observed Measures With the Sample Size at All Timepoints for Grade 5

Measure	1	2	3	4	5	6	7	8	9	10
1. ORF 1	462									
2. ORF 2	**.92**	461								
3. ORF 3	**.92**	**.93**	461							
4. ORF 4	**.88**	**.88**	**.89**	482						
5. ORF 5	**.88**	**.89**	**.89**	**.93**	482					
6. ORF 6	**.88**	**.90**	**.88**	**.92**	**.94**	482				
7. ORF 7	**.88**	**.87**	**.89**	**.91**	**.91**	**.90**	443			
8. ORF 8	**.87**	**.87**	**.88**	**.91**	**.91**	**.91**	**.95**	443		
9. ORF 9	**.89**	**.90**	**.90**	**.92**	**.92**	**.92**	**.94**	**.94**	443	
10. NTRP	.55	.54	.55	.49	.53	.53	.49	.51	.54	302

Note. NTRP = national tests of reading proficiency; ORF = oral reading fluency measure. The bold numbers are correlations within construct. All correlations are significant at the .01 level. The diagonal is the sample size for each measure.

occasions) and concurrent relations with it (spring). Relations between the ORF measure and the simple chains of words subtest (r = .73, .75, and .75, respectively) were stronger than with the three other subtests that targeted higher level reading skills (reading narrative text: r = .49, .53, and .51, respectively; reading expository text: r = .44, .47, and .47, respectively) and vocabulary (word knowledge: r = .31, .36, and .35, respectively).

The fifth-grade national reading test was completed by grade 5 students at the time of the ORF measure's administration in the fall (concurrent relation) and by grade 4 students in the fall of the year after the ORF measure's administration when the students had moved to grade 5 (predictive relation). The concurrent and predictive relations between the ORF measure and the two fifth-grade samples on the national test are estimated to be .55 and .54, respectively. These somewhat lower correlations are in line with expectations, given that the national reading test focuses on higher level reading competences, such as finding information, reading comprehension, text interpretation, and reflection. An increase in ORF scores during grade 5 occurs, but the rate of acceleration is flatter.

Discussion

The main purpose of the study was to examine initial status and growth on ORF and how ORF relates to students' reading performance in and across multiple grades on two high-stakes national compulsory reading tests focused on decoding and comprehension in Norwegian, a semitransparent orthography. The study differed from previous research in that it examined the use of an ORF measure in a European context with a more transparent language than English using longitudinal second-order latent growth curve modeling. Overall, this study makes several contributions to the existing research regarding growth trajectories on the ORF measure and its relations with general reading proficiency.

First, we examined whether the ORF measure has longitudinal measurement invariance. This is important because invariance in text difficulty and complexity can help determine that the growth trajectories are due to the reading development and not passage characteristics. We found that the configural reference model for the ORF passages was measuring the same underlying construct. This was predicted because all passages were constructed to be parallel items, similar in difficulty but with different content (stories) to avoid retest effects. However, when the intercepts of the ORF scores were restricted to be equal across time, particular passages stood out empirically as more or less difficult and affected the general trend in ORF latent means. By relaxing some of the restrictions for these reading passages, a well-fitting partial scalar invariance model could still be obtained for each grade level (2–5). This allowed for meaningful, unambiguous comparisons and further analyses of the ORF measure across time. We explored potential reasons for why some passages differed empirically from the others by linking the deviations in intercept (mean) and loading of the passages to (a) technical measures of readability (e.g., LIX readability formula: Gilliland, 1972; Flesch readability formula: Flesch & Paterson, 1948) and (b) the content type and topic of the reading passage (see Appendix A). However, we did not find a link with the technical measures nor with content type.

The invariance results support previous research and show the difficulty of creating parallel reading passages (e.g., Cummings, Park, & Bauer Schaper, 2013). Although the grade-level passages are developed to be equal in difficulty, the reason why some passages stood out empirically might be that different passages mirror students' interests and familiarity because of content variation in the passages and the types of text structures used (e.g., narrative, expository). Our findings underline the importance of measuring ORF across a set of reading passages instead of basing results on only a one-passage measure. In fact, an observed median score of the three passages at each measurement point is recognized as a better indicator of a student's ORF performance than just one passage (DIBELS: Good & Kaminski, 2002).

Second, we examined students' growth in ORF. As for the examination of growth trajectories in a semitransparent language, the main findings—that linear growth represented grades 2 through 3, and more nonlinear growth represented grades 4 and 5—were as expected and extend previous research in students' growth of reading fluency. The findings that the Norwegian students' performance increased over the course of the year and fastest in grades 2 and 3 are similar to previous studies (see, e.g., Hasbrouck & Tindal, 1992, 2006). The slower growth in grade 5 might be interpreted as an indication of reaching a performance ceiling, especially for some students, and hence a flattening out in level of ORF rate in Norwegian. However, further studies including students in higher grades should be conducted to determine this.

The nonlinear growth pattern is supported by previous research of ORF growth in English for students in the later grades (e.g., S.K. Baker et al., 2008; Fuchs et al., 1993; Nese et al., 2013). Furthermore, it is consistent with theory and research regarding the development of automaticity in reading (LaBerge & Samuels, 1974). The nonlinear growth pattern indicates that as reading is first developing, changes in fluency are reflecting that the decoding process is becoming more automatic. As students become more proficient in reading, individual differences seem to deal more with reading comprehension of the particular passage than with reading comprehension in general (e.g., García & Cain, 2014; Pikulski & Chard, 2005; Stanovich, 2000). This does not necessarily mean that administering the ORF measure in grade 5 is not useful but that expectations for linear growth might be unrealistic in practice (Nese et al., 2013).

The context for this study is that reading fluency was measured in a semitransparent orthography, whereas most other studies of ORF measure reading fluency in English, which has a less transparent orthography. Previous studies comparing the development of reading fluency between students from different orthographies have found that there are similar mechanisms and predictors underlying the development of decoding but that students learning to read in English have slower decoding growth rates, at least during the first three years of school (Caravolas et al., 2013). If we compare our findings with growth rates found in studies of English readers, we see that the Norwegian students began the year reading fewer WCPM but experience stronger growth during the school year. For instance, Tindal, Nese, Stevens, and Alonzo (2015) found that U.S. students in grade 3 began the year reading an average of nearly 81 WCPM (15 WCPM more than Norwegian students), just above 100 WCPM in grade 4 (three WCPM more than Norwegians), and 125 WCPM in grade 5 (21 WCPM more than Norwegians).

The slope in each of these grades ranged between 0.65 and 0.73 WCPM, whereas the slope in the Norwegian grades 3–5 ranged between 0.76 and 0.84 WCPM.

The lower initial value of WCPM in Norwegian students might be due to the fact that learning to read starts at a later age than for students in the U.S. school system. However, the stronger ORF growth in Norwegian students supports previous research findings that learning to read in a more transparent language is easier, particularly in the early stages of reading development, than learning to read in a nontransparent orthography (Caravolas et al., 2013). Another possible explanation is that reading instruction during the school year in Norwegian elementary schools is somehow different from other contexts and more aligned with practices that accelerate reading fluency growth. Another important factor is that U.S. students start their formal reading instruction one year earlier than Norwegian students.

As expected, individual differences in initial ORF level varied much more than growth rates. However, all individuals had positive growth rates. The estimated population variation in growth across individuals was largest in grades 3 and 4, and the correlation between initial level and growth in these grades was small. In grades 2 and 5, the variance in growth rates across individuals was small and almost nonexistent. One interpretation is that the effects of reading instruction are constant for students across different reading-proficiency levels and that the students are relatively homogeneous in terms of their reading proficiency. The initial differences might be useful as a baseline indicator to identify students at risk for reading problems, which was demonstrated in several previous studies (e.g., Silberglitt & Hintze, 2007; Speece & Ritchey, 2005; Wang, Porfeli, & Algozzine, 2008).

Based on early screening to identify struggling readers, Parrila et al. (2005) demonstrated that it is possible for teachers to reduce individual differences in basic reading skills during early reading development. Teachers can respond early to individual differences among students with specific interventions, followed by systematic monitoring of students' growth (Stecker, Fuchs, & Fuchs, 2005). Although variability in ORF growth should be expected, results of the present study and others can be used to define normative rates of growth that can help identify students with low initial ORF levels and/or slow ORF growth so they can receive more intensive reading support. Furthermore, it will be important to define ORF benchmark and/or cutoff scores to identify struggling readers for intervention and progress monitoring. However, thresholds for appropriate levels of automaticity and reading rate by grade level might best be considered by using receiver operating characteristic curves using generated specifications related to sensitivity and specificity. Establishing thresholds using professional judgment and various objective approaches (e.g., students scoring below the 20th percentile) might help teachers identify the "right" students for intervention or extra support, but the sensitivity of yielding "true positives" and the specificity of yielding "true negatives" is also important to consider (Smolkowski, Cummings, & Strycker, 2016).

Finally, we investigated the relative stability of ORF growth and the concurrent and predictive relations between the ORF measure and the NTRP. The high stability of the ORF measure, reflected by correlations between the three ORF factors at all three timepoints in all four grade levels (>.92) in a semitransparent orthography, confirms previous evidence generated in less transparent orthographies (e.g., S.K. Baker et al., 2008; Kim et al., 2010; Nese et al., 2013). Overall, the findings extend evidence to more transparent languages by demonstrating moderate to strong positive correlations between the ORF measure and the NTRP across the school year in grades 2–5 (range = .44–.75). The findings are in line with previous studies on correlations between the ORF measure and high-stakes criterion measures of reading in English. However, high-stakes criterion measures among state tests and national tests have varying levels of difficulty and psychometric quality, of course. For instance, Wanzek et al. (2010) demonstrated in a longitudinal study of predictive validity across grades 1–3 in the United States that the ORF measure was a reliable predictor of students' reading proficiency on two different high-stakes measures in grade 3. However, greater student growth on the ORF measure through the three grades was needed to achieve success on the nationally normed test (SAT-10) compared with what was needed on the state-normed test (Texas Assessment of Knowledge and Skills).

Furthermore, in a study across grades 1–3, S.K. Baker et al. (2008) found slightly different correlations between the ORF measure and two high-stakes measures, the SAT-10 (range = .63–.80) and the Oregon state test (range = .58–.68). Findings that the ORF measure provided a stronger relation to the NTRP in earlier grades than in later grades in Norway have also been confirmed by previous research on how the ORF measure relates to high-stakes criterion measures in reading in a less transparent orthography and how relations between ORF and reading performance decrease over time (e.g., S.K. Baker et al., 2008).

Regarding implications for practice, our study showed that the ORF measure is an important developmental indicator of reading proficiency and is useful in monitoring students' reading fluency, which can help schools identify students who are at risk for reading failure (Fuchs et al., 2001; Pfost et al., 2012). In a prevention and early intervention framework of reading development, ORF is an efficient measure that schools can use to help teachers efficiently identify students who are on track and those who are not. This can lead to providing struggling readers with targeted support in the early stages of reading development, when their growth trajectories in some areas, such as reading fluency, tend to be developing rapidly (S.K. Baker et al., 2008; Hosp & Suchey, 2014; Pikulski & Chard, 2005). By identifying struggling readers early and following their development within the year, teachers can initiate early reading supports and do not have to wait for high-stakes tests, such as the NTRP results, at the end of the school year. It is worth noting that the absence of the NTRP in Norwegian in grade 4 probably increases the risk of not identifying struggling readers. That is, two years can elapse before reading data are provided. In summary, we conclude that the Norwegian version of the ORF measure is a reliable and valid screening instrument that is easy and efficient to administer in schools and contributes to the early identification of students at risk for reading difficulties across years in the elementary grades.

Because poor reading skills can be a significant impediment to success in formal education, interventions are crucial. As shown in the present study, and also demonstrated in previous studies, the ORF measure can serve as an index of students' reading development, not only as a measure of reading fluency per se. Many studies have shown that reading fluency problems can be effectively remediated through repeated reading interventions (for a review, see Chard, Vaughn, & Tyler, 2002; National Institute of Child Health and Human Development, 2000). However, these kinds of repeated reading interventions can lead to reading instruction practices where the focus is on reading for speed and where other important components are excluded (Rasinski et al., 2011; Rayner, Schotter, Masson, Potter, & Treiman, 2016). Thus, because the ORF measure also serves as an index of reading problems beyond reading fluency, repeated reading should not be the only intervention for these students. Many studies have shown that the best way to ensure strong comprehension and with a sufficient reading speed is to also work on vocabulary, in line with the simple view of reading. However, learning to read is a complex process that includes several aspects beyond the simple view (e.g., sociocultural, neurological, genetic). For instance, from a sociocultural point of view, Purcell-Gates (2002) argued that the simple view of reading is not without controversy. Because students enter different school contexts from different socioeconomic backgrounds, they will face different learning and reading difficulties. Ultimately, it is important not only to identify each student's specific needs by analyzing potential needs from different perspectives but also to differentiate the interventions appropriately so all students are supported effectively.

Also, substantial research has shown that for young students struggling with learning to read, small-group interventions that emphasize all major aspects of reading development (phonics, fluency, comprehension, and vocabulary) consistently produce benefits in measured aspects of reading, including comprehension and fluency (Hulme & Melby-Lervåg, 2015; Melby-Lervåg & Lervåg 2014a). Furthermore, even for students with fluency problems solely, interventions could focus on fluency, including instructions for comprehension and prosody. Notably, students with dyslexia with no additional language problems will also get a low score on the ORF measure but will not necessarily need vocabulary and language comprehension training as a part of their intervention. It is therefore important to monitor progress and provide more in-depth diagnostic assessments to determine in a more precise way what specific areas of difficulty a student is having trouble with. Thus, teachers can distinguish between using the ORF measure for screening and progress-monitoring assessments and when additional assessment data are needed for other purposes, such as determining specific program areas and intervention content. Furthermore, additional diagnostic data will be necessary when more intensive interventions for students are needed because they are not responding sufficiently to universal or less intense interventions (e.g., vocabulary, decoding).

Future Studies

In future studies, it could be useful to validate the ORF measure in a transparent orthography against a larger battery of diagnostic reading tests in addition to more general national or statewide assessment tests. By validating the ORF measure against a battery of individually administered reading tests, it would be possible to examine how sensitive and specific the ORF test is when it comes to detecting reading problems at an early stage (see Duff et al., 2015; Snowling & Hulme, 2012). The ORF measure might also be validated against other groups of students with known characteristics (e.g., dyslexia, individualized education plans in reading). Although we did not find significant differences on the ORF measure by student group based on variables such as gender or in the small group of bilingual students in this study, these variables in addition to students' socioeconomic status will be important to investigate more thoroughly in future studies.

The ORF measure as implemented in this study was a teacher-administered measure, which can affect data collection quality based on teachers' prior knowledge of students, potential bias against or in favor of the measure or specific students, or potential bias in relation to the use of test data (e.g., accountability, teacher evaluations). Thus, the ORF measure should be validated against tests that are not administered or collected by teachers, or when possible, ORF data should be collected by impartial examiners. However, because the ORF measure is intended as a teacher instrument, it is important to also have teachers as test administrators. Still, in future studies, inter-rater reliabilities should be included. Finally, in both Europe and the United States, large-scale full-classroom assessments have also been met by considerable controversy among educators (e.g., Goodman, 2006; National Union of Teachers, 2012). Hence, usability and usefulness of the ORF measure from a teacher perspective needs to be evaluated. To our knowledge, this type of study has not yet been conducted. In contrast to the mandatory tests used in schools, the ORF measure is not a one-time check but a tool that can be integrated into teachers' work throughout the year to measure students' progress. This is potentially more useful because it can be more directly linked to intervention and used as a measure of progress.

NOTE

We gratefully acknowledge the students and their teachers for participating in the study.

REFERENCES

Baker, D.L., Stoolmiller, M., Good, R., & Baker, S. (2011). Effect of reading comprehension on passage fluency in Spanish and English for second-grade English learners. *School Psychology Review*, *40*(3), 331–351.

Baker, S.K., Smolkowski, K., Katz, R., Fien, H., Seeley, J.R., Kame'enui, E.J., & Beck, C.T. (2008). Reading fluency as a predictor of reading proficiency in low-performing, high-poverty schools. *School Psychology Review*, *37*(1), 18–37.

Baker, S.K., Smolkowski, K., Smith, J.M., Fien, H., Kame'enui, E.J., & Beck, C.T. (2011). The impact of Oregon Reading First on student reading outcomes. *The Elementary School Journal*, *112*(2), 307–331. doi:10.1086/661995

Breznitz, A. (2006). *Fluency in reading: Synchronization of processes*. Mahwah, NJ: Erlbaum.

Caravolas, M., Lervåg, A., Defior, S., Seidlová Málková, G., & Hulme, C. (2013). Different patterns, but equivalent predictors, of growth in reading in consistent and inconsistent orthographies. *Psychological Science*, *24*(8), 1398–1407. doi:10.1177/0956797612473122

Caravolas, M., Lervåg, A., Mousikou, P., Efrim, C., Litavsky, M., Onochie-Quintanilla, E., … Hulme, C. (2012). Common patterns of prediction of literacy development in different alphabetic orthographies. *Psychological Science*, *23*(6), 678–686. doi:10.1177/0956797611434536

Chard, D.J., Vaughn, S., & Tyler, B.-J. (2002). A synthesis of research on effective interventions for building fluency with elementary students with learning disabilities. *Journal of Learning Disabilities*, *35*(5), 386–406. doi:10.1177/00222194020350050101

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233–255. doi:10.1207/S15328007SEM0902_5

Cromley, J.G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*(2), 311–325. doi:10.1037/0022-0663.99.2.311

Cummings, K.D., Biancarosa, G., Schaper, A., & Reed, D.K. (2014). Examiner error in curriculum-based measurement of oral reading. *Journal of School Psychology*, *52*(4), 361–375. doi:10.1016/j.jsp.2014.05.007

Cummings, K.D., Park, Y., & Bauer Schaper, H.A. (2013). Form effects on DIBELS Next Oral Reading Fluency progress-monitoring passages. *Assessment for Effective Intervention*, *38*(2), 91–104. doi:10.1177/1534508412447010

Deno, S.L., Mirkin, P.K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, *49*(1), 36–45.

Duff, F.J., Mengoni, S.E., Bailey, A.M., & Snowling, M.J. (2015). Validity and sensitivity of the phonics screening check: Implications for practice. *Journal of Research in Reading*, *38*(2), 109–123. doi:10.1111/1467-9817.12029

Flesch, R., & Paterson, D.G. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. doi:10.1037/h0057532

Foorman, B.R., Koon, S., Petscher, Y., Mitchell, A., & Truckenmiller, A. (2015). Examining general and specific factors in the dimensionality of oral language and reading in 4th–10th grades. *Journal of Educational Psychology*, *107*(3), 884–899. doi:10.1037/edu0000026

Fuchs, L.S., Fuchs, D., Hamlett, C.L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review*, *22*(1), 27–48.

Fuchs, L.S., Fuchs, D., Hosp, M.K., & Jenkins, J.R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, *5*(3), 239–256. doi:10.1207/S1532799XSSR0503_3

García, J.R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research*, *84*(1), 74–111. doi:10.3102/0034654313499616

García-Madruga, J.A., Vila, J.O., Gómez-Veiga, I., Duque, G., & Elosúa, M.R. (2014). Executive processes, reading comprehension and academic achievement in 3th grade primary students. *Learning and Individual Differences*, *35*, 41–48. doi:10.1016/j.lindif.2014.07.013

Gilliland, J. (1972). *Readability*. London, UK: University of London Press.

Good, R.H., & Kaminski, R.A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Education Achievement.

Good, R.H., Kaminski, R.A., & Dill, S. (2002). DIBELS oral reading fluency and retell fluency. In R.H. Good & R.A. Kaminski (Eds.), *Dynamic Indicators of Basic Early Literacy Skills* (6th ed., pp. 30–38). Eugene, OR: Institute for the Development of Education Achievement.

Goodman, K.S. (with Flurkey, A., Kato, T., Kamii, C., Manning, M., Seay, S., Thome, C., … Wilde, S.). (2006). *The truth about DIBELS: What it is, what it does*. Portsmouth, NH: Heinemann.

Gough, P., & Tunmer, W. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*(1), 6–10. doi:10.1177/074193258600700104

Gustafsson, J.E., Allodi Westling, M., Alin Åkerman, B., Eriksson, C., Eriksson, L., Fischbein, S., … Persson, R.S. (2010). *School, learning and mental health: A systematic review*. Stockholm, Sweden: The Royal Swedish Academy of Sciences, The Health Committee.

Hasbrouck, J., & Tindal, G. (1992). Curriculum-based oral reading fluency norms for students in grades 2 through 5. *Teaching Exceptional Children*, *24*(3), 41–44. doi:10.1177/004005999202400310

Hasbrouck, J., & Tindal, G. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher*, *59*(7), 636–644. doi:10.1598/RT.59.7.3

Hoover, W.A., & Gough, P.B. (1990). The simple view of reading. *Reading and Writing*, *2*(2), 127–160. doi:10.1007/BF00401799

Horn, J.L., & McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144. doi:10.1080/03610739208253916

Hosp, J., & Suchey, N. (2014). Reading assessment: Reading fluency, reading fluently, and comprehension—Commentary on the special topic. *School Psychology Review*, *43*(1), 59–68.

Hu, L.T., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. doi:10.1080/10705519909540118

Hulme, C., Bowyer-Crane, C., Carroll, J.M., Duff, F.J., & Snowling, M.J. (2012). The causal role of phoneme awareness and letter-sound knowledge in learning to read. *Psychological Science*, *23*(6), 572–577. doi:10.1177/0956797611435921

Hulme, C., & Melby-Lervåg, M. (2015). Effects from interventions for psychological learning and behavioural disorders in children. In A. Thapar, D.S. Pine, J.F. Leckman, S. Scott, M.J. Snowling, & E. Taylor (Eds.), *Rutter's child and adolescent psychiatry* (6th ed., pp. 533–545). Oxford, UK: John Wiley & Sons.

Kim, Y.S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology*, *102*(3), 652–667. doi:10.1037/a0019643

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction–integration model. *Psychological Review*, *95*(2), 163–182. doi:10.1037/0033-295X.95.2.163

Kuhn, M.R., & Stahl, S.A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, *95*(1), 3–21. doi:10.1037/0022-0663.95.1.3

LaBerge, D., & Samuels, S.A. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*(2), 293–323. doi:10.1016/0010-0285(74)90015-2

Landerl, K., & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology*, *100*(1), 150–161. doi:10.1037/0022-0663.100.1.150

Lervåg, A., & Aukrust, V.G. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *The Journal of Child Psychology and Psychiatry*, *51*(5), 612–620. doi:10.1111/j.1469-7610.2009.02185.x

Little, T.D. (2013). *Longitudinal structural equation modeling.* New York, NY: Guilford.

Melby-Lervåg, M., & Lervåg, A. (2014a). Effects from educational interventions on reading comprehension and its underlying components. *Child Development Perspectives*, *8*(2), 96–100. doi:10.1111/cdep.12068

Melby-Lervåg, M., & Lervåg, A. (2014b). Reading comprehension and its underlying components in second-language learners: A meta-analysis of studies comparing first- and second-language learners. *Psychological Bulletin*, *140*(2), 409–433. doi:10.1037/a0033890

Millsap, R.E. (2011). *Statistical approaches to measurement invariance.* New York, NY: Routledge.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction, reports of the subgroups* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

National Union of Teachers. (2012). *'Five years old is too young to fail!': Year one phonics screening check report.* Retrieved from https://www.teachers.org.uk/files/phonics-survey-conference-2012-apr-12-je.doc

Nese, J.F.T., Biancarosa, G., Cummings, K., Kennedy, P., Alonzo, J., & Tindal, G. (2013). In search of average growth: Describing within-year oral reading fluency growth across grades 1–8. *Journal of School Psychology*, *51*(5), 625–642. doi:10.1016/j.jsp.2013.05.006

Norwegian Reading Centre, University of Stavanger. (2013a). *Kartleggingsprøve i lesing på 2.trinn—forslag til endelig prøve* [Mandatory reading assessment in grade 2—proposed final test; Technical report]. Stavanger, Norway: Author.

Norwegian Reading Centre, University of Stavanger. (2013b). *Kartleggingsprøve i lesing på 3.trinn—forslag til endelig prøve* [Mandatory reading assessment in grade 3—proposed final test; Technical report]. Stavanger, Norway: Author.

Organisation for Economic Co-operation and Development (2013). *OECD economic surveys: France 2013.* Paris, France: Author.

Parrila, R., Aunola, K., Leskinen, E., Nurmi, J.-E., & Kirby, J.R. (2005). Development of individual differences in reading: Results from longitudinal studies in English and Finnish. *Journal of Educational Psychology*, *97*(3), 299–319. doi:10.1037/0022-0663.97.3.299

Perfetti, C. (1985). *Reading ability.* New York, NY: Oxford University Press.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22–37. doi:10.1080/10888438.2013.827687

Pfost, M., Dörfler, T., & Artelt, C. (2012). Reading competence development of poor readers in a German elementary school sample: An empirical examination of the Matthew effect model. *Journal of Research in Reading*, *35*(4), 411–426. doi:10.1111/j.1467-9817.2010.01478.x

Pikulski, J.J., & Chard, D.J. (2005). Fluency: Bridge between decoding and reading comprehension. *The Reading Teacher*, *58*(6), 510–519. doi:10.1598/RT.58.6.2

Purcell-Gates, V. (2002). The irrelevancy—and danger—of the 'simple view' of reading to meaningful standards. In R. Fisher, M. Lewis, & G. Brooks (Eds.), *Raising standards in literacy* (pp. 105–116). London, UK: RoutledgeFalmer.

Rasinski, T.V., Reutzel, C.R., Chard, D., & Linan-Thompson, S. (2011). Reading fluency. In M.L. Kamil, P.D. Pearson, E.B. Moje, & P. Afflerbach (Eds.), *Handbook of reading research* (Vol. 4, pp. 286–319). New York, NY: Routledge.

Rayner, K., Schotter, E.R., Masson, M.E.J., Potter, M.C., & Treiman, R. (2016). So much to read, so little time: How do we read, and can speed reading help? *Psychological Science in the Public Interest*, *17*(1), 4–34. doi:10.1177/1529100615623267

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*(3), 726–748. doi:10.1037/0033-2909.92.3.726

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi:10.18637/jss.v048.i02

Schwanenflugel, P.J., Hamilton, A.M., Kuhn, M.R., Wisenbaker, J., & Stahl, S.A. (2004). Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers. *Journal of Educational Psychology*, *96*(1), 119–129. doi:10.1037/0022-0663.96.1.119

Seidenberg, M.S., & McClelland, J.L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568. doi:10.1037/0033-295X.96.4.523

Shinn, M.R. (Ed.). (1989). *Curriculum-based measurement: Assessing special children*. New York, NY: Guilford.

Shinn, M.R. (1998). *Advanced applications of curriculum-based measurement*. New York, NY: Guilford.

Shinn, M.R., Shinn, M.M., Hamilton, C., & Clarke, B. (2002). Using curriculum-based measurement in general education classrooms to promote reading success. In M.R. Shinn, H.M. Walker, & G. Stoner (Eds.), *Interventions for academic and behavior problems*

II: Prevention and remedial approaches (pp. 113–142). Bethesda, MD: National Association of School Psychologists.

Silberglitt, B., & Hintze, J.M. (2007). How much growth can we expect? A conditional analysis of R-CBM growth rates by level of performance. *Exceptional Children*, *74*(1), 71–84. doi:10.1177/001440290707400104

Skaftun, A., Stangeland, E.B., Solheim, O.J., & Mangen, A. (2013). *Den nasjonale prøven i lesing på 5.trinn, 2013* [The national reading test in grade 5, 2013; Technical report]. Stavanger, Norway: The Norwegian Reading Centre, University of Stavanger.

Smolkowski, K., Cummings, K.D., & Strycker, L. (2016). An introduction to the statistical evaluation of fluency measures with signal detection theory. In K.D. Cummings, & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and application* (Vol. 1, pp. 187–221). New York, NY: Springer.

Snowling, M.J., & Hulme, C. (2012). Annual research review: The nature and classification of reading disorders—a commentary on proposals for DSM–5. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *53*(5), 593–607.

Solheim, O.J., Skaftun, A., & Walgermo, B.R. (2012). *Den nasjonale prøven i lesing på 5.trinn, 2012* [The national reading test in grade 5, 2012; Technical report]. Stavanger, Norway: The Norwegian Reading Centre, University of Stavanger.

Speece, D.L., & Ritchey, K.D. (2005). A longitudinal study of the development of oral reading fluency in young children at risk for reading failure. *Journal of Learning Disabilities*, *38*(5), 387–399. doi:10.1177/00222194050380050201

Stage, S., & Jacobsen, M. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review*, *30*(3), 407–419.

Standards & Testing Agency. (2016). *National curriculum assessments: Past papers*. Retrieved from https://www.gov.uk/government/collections/key-stage-2-tests-past-papers#phonics-screening-check

Stanovich, K.E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York, NY: Guilford.

Statens Beredning för Medicinsk Utvärdering (2014). *Dyslexi hos barn och ungdomar—tester och innsatser: En systematisk litteraturöversikt* [Dyslexia in children and adolescence—tests and efforts: A systematic review]. Stockholm, Sweden: Author.

Stecker, P.M., Fuchs, L.S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools*, *42*(8), 795–819. doi:10.1002/pits.20113

Stoolmiller, M. (1995). Using latent growth curve models to study developmental processes. In J.M. Gottman (Ed.), *The analysis of change* (pp. 103–138). Mahwah, NJ: Erlbaum.

Stoolmiller, M., Biancarosa, G., & Fien, H. (2013). Measurement properties of DIBELS Oral Reading Fluency in grade 2: Implications for equating studies. *Assessment for Effective Intervention*, *38*(2), 76–90. doi:10.1177/1534508412456729

Tindal, G., Nese, J.F.T., Stevens, J.J., & Alonzo, J. (2015). Growth on oral reading fluency measures as a function of special education and measurement sufficiency. *Remedial and Special Education*, *37*(1), 28–40. doi:10.1177/0741932515590234

van IJzendoorn, M.H., & Bus, A.G. (1994). Meta-analytic confirmation of the nonword reading deficit in developmental dyslexia. *Reading Research Quarterly*, *29*(3), 266–275. doi:10.2307/747877

Veenendaal, N.J., Groen, M.A., & Verhoeven, L. (2015). What oral text reading fluency can reveal about reading comprehension. *Journal of Research in Reading*, *38*(3), 213–225. doi:10.1111/1467-9817.12024

Wang, C., Porfeli, E., & Algozzine, B. (2008). Development of oral reading fluency in young children at risk for failure. *Journal of Education for Students Placed at Risk*, *13*(4), 402–425. doi:10.1080/10824660802427702

Wanzek, J., Roberts, G., Linan-Thompson, S., Vaughn, S., Woodruff, A.L., & Murray, C.S. (2010). Differences in the relationship of oral reading fluency and high-stakes measures of reading comprehension. *Assessment for Effective Intervention*, *35*(2), 67–77.

Widaman, K.F., Ferrer, E., & Conger, R.D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10–18. doi:10.1111/j.1750-8606.2009.00110.x

Wise, J.C., Sevcik, R.A., Morris, R.D., Lovett, M.W., Wolf, M., Kuhn, M., … Schwanenflugel, P. (2010). The relationship between different measures of oral reading fluency and reading comprehension in second-grade students who evidence different oral reading fluency difficulties. *Language, Speech, and Hearing Services in Schools*, *41*(3), 340–348. doi:10.1044/0161-1461(2009/08-0093)

Submitted February 11, 2016
Final revision received June 24, 2016
Accepted June 27, 2016

ANNE ARNESEN (corresponding author) is a doctoral student in the Department of Special Needs Education at the University of Oslo, Norway; e-mail anne.arnesen@isp.uio.no. She is interested in aspects of evidence-based assessments for early identification of children at risk for difficulties in reading and social behavior and how Response to Intervention can promote their development and learning.

JOHAN BRAEKEN is an associate professor in the Centre for Educational Measurement at the University of Oslo, Norway; e-mail johan.braeken@cemo.uio.no.

SCOTT BAKER is a research professor at the Center on Research and Evaluation at the Simmons School of Education and Human Development, Southern Methodist University, Dallas, Texas, USA; e-mail skbaker@smu.edu. He is interested in the impact of interventions on child outcomes, mechanisms that underlie effective interventions, and how intervention impact varies by factors intrinsic and extrinsic to the child.

WILHELM MEEK-HANSEN is a special advisor at the Norwegian Center for Child Behavioral Development, Oslo, Norway; e-mail wilhelm.meek-hansen@atferdssenteret.no. He is interested in the relationship between children's emotional and behavioral problems and reading difficulties that impact their social and academic development.

TERJE OGDEN is the research director at the Norwegian Center for Child Behavioral Development, Oslo, Norway; e-mail terje.ogden@atferdssenteret.no.

MONICA MELBY-LERVÅG is a professor in the Department of Special Needs Education at the University of Oslo, Norway; e-mail monica.melby-lervag@isp.uio.no. She is interested in language and reading development in children with dyslexia, specific language difficulties, and minority languages.

APPENDIX A

Textual Properties of the Reading Passages Related to Empirical Performance Deviations From the Intended Parallel Construction

	ΔDIF[a]		LIX readability formula		Flesch readability formula		Summary	Content	
Reading passage	Intercept	Loading	Full	Average read	Reading ease	Grade level	Grade level	Topic	Type
Grade 2									
1	0.00	1.00	11	20	92.9	2.2	2	Dear diary	Narrative
2	3.13***	1.00	12	9	96.2	2.4	2	Ill	Expository
3	0.58	0.90***	13	8	96.6	2.2	2	Pen friend	Narrative
4	0.00	1.04***	12	10	92.3	2.4	2	Little and nice	Expository
5	0.00	1.00	11	9	97.2	1.7	2	In the cabin	Narrative
6	0.00	1.00	15	11	88.8	3.2	3	On way to school	Expository
7	6.24***	1.00	10	7	97.5	2.3	2	Afraid of darkness	Narrative
8	0.00	1.00	12	11	98.6	1.8	2	Angry	Expository
9	0.00	1.00	13	9	93.3	2.3	2	Life along river	Expository
Grade 3									
10	0.00	1.00	14	16	93.9	2.2	3	Soccer tournament	Narrative
11	5.81***	1.00	17	13	86.1	3.8	4	In the library	Expository
12	0.00	1.00	19	20	87.2	3.3	4	School camp	Narrative
13	0.00	1.00	18	14	85.8	3.9	4	Boys don’t play	Expository
14	0.00	1.00	18	18	87.0	3.3	4	Grandfather fishing	Narrative
15	0.00	1.00	18	19	85.4	4.0	4	Wild animals	Narrative
16	0.00	1.00	15	12	90.5	3.0	3	Stranger	Narrative
17	−6.94***	1.00	17	11	87.9	3.4	4	An iceland is born	Expository
18	0.00	1.00	18	17	86.6	3.6	4	Moving to town	Narrative
Grade 4									
19	0.00	1.00	20	19	81.6	4.5	5	Author visiting	Narrative
20	12.40***	1.00	20	19	82.4	4.6	4	Moving to London	Narrative
21	0.00	1.00	21	22	83.0	4.6	5	Kayaking	Expository
22	0.00	1.00	21	27	80.8	4.5	5	Bike ride	Narrative
23	6.83***	1.14***	20	18	85.3	4.8	4	Dog in the house	Expository
24	0.00	1.01*	19	18	80.7	4.4	5	Cheeta (cat)	Expository
25	0.00	1.00	23	22	81.1	4.8	5	Water well	Narrative
26	4.97***	1.00	20	24	79.4	5.1	4	Emil and Eilert	Narrative
27	−10.52***	1.00	22	21	83.2	4.7	5	Cairo	Expository

(*continued*)

Textual Properties of the Reading Passages Related to Empirical Performance Deviations From the Intended Parallel Construction (*continued*)

	ΔDIF[a]		LIX readability formula		Flesch readability formula		Summary	Content	
Reading passage	Intercept	Loading	Full	Average read	Reading ease	Grade level	Grade level	Topic	Type
Grade 5									
28	0.00	1.02***	22	23	81.9	3.9	5	Soccer game	Narrative
29	0.00	1.00	24	29	77.0	5.3	6	Vikings	Expository
30	15.36***	1.00	22	27	78.3	5.7	6	Save the children	Expository
31	−10.63***	1.00	26	26	75.2	5.8	5	Way guide	Narrative
32	−17.28***	1.12***	25	22	77.4	5.0	6	Climbing wall	Narrative
33	0.00	1.00	27	23	77.2	5.1	5	Dolphins	Expository
34	4.87***	1.00	19	23	76.9	5.0	5	Family trip	Narrative
35	0.00	1.00	25	25	77.0	5.5	6	Water is source	Expository
36	0.00	1.00	26	30	73.6	5.9	6	The body is not	Expository

[a] Wald tests were run for the ΔDIF parameters.
$^{*}p < .05$. $^{***}p < .001$.

APPENDIX B

Subtest	Number of items	Timed	Description
National reading assessment subtests for grade 2: Spring 2013			
Recognizing letters	25	1 minute	The students were presented printed capital letters and asked to identify the same printed lowercase letters and vice versa.
Writing words	16	—	To measure skills in spelling, the examiner asked the students to write words when listening to the teacher read words aloud.
Reading words	21	2 minutes	The students were asked to look at a picture and identify which of four words represented the picture.
Splitting compound words	21	1 minute	To measure morphemic awareness and word-decoding skill, the students were asked to divide compound words by putting a line between two meaningful units in words that varied in terms of the number of letters and difficulty.
Reading sentences	18	2 minutes	To measure reading comprehension at the sentence level, the students were asked to read a sentence of increasing length (two to eight words) and identify which of four pictures best represented the meaning of the sentence.
Following written instructions	10	2.5 minutes	The students were asked to read instructions (one or two sentences) and demonstrate their reading comprehension by marking on a picture of elements the one that corresponded to each instruction (e.g., “Please make a circle around the bus station”).
Reading text	6	—	To measure reading comprehension, the students read one short text silently and then answered six multiple-choice questions about the text, which was taken from Aesop’s fable “The Bear and the Two Friends.”

(*continued*)

(***continued***)

Subtest	Number of items	Timed	Description
National reading assessment subtests for grade 3: Spring 2013			
Chains of words	66	5 minutes	To measure decoding and word recognition skills, the students were presented an unbroken chain of four meaningful words (e.g., “onfivebeatcheese”) and asked to read the unit as separate words (“*on five beat cheese*”).
Reading narrative text	9	15 minutes	To measure reading comprehension, this subtest consisted of a narrative text that students read silently and then answered nine questions about it. The text was from a Norwegian illustrated children’s book. Three questions measured literacy comprehension, in that the students could find the information to answer the questions in the text. Five other questions measured inferential comprehension, in that the students had to integrate information from the text with their own background knowledge about the topic or infer meaning in the text from things not stated explicitly in the text to answer the question correctly; one question required students to make a reasonable interpretation of the text based on multiple pieces of information in the text.
Word knowledge	20	—	Multiple-choice items measured vocabulary. Each item consisted of four words. The teacher read the first target word aloud and then each of three option words, one of which was a synonym for the target word. The students marked the word that was the correct synonym.
Reading expository text	7	15 minutes	To measure reading comprehension, the students silently read a text about making pancakes and answered seven multiple-choice questions about the text. Information for five of the questions was directly expressed in the text. For the other two questions, the students had to combine information from different places in the text and rely on their own background knowledge and experience to answer the questions correctly.