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, Hamilton, & Clarke, 2002).

In Europe, however, few systematic studies have been conducted concerning the instruments that schools use to assess reading skills.
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, & Tunmer, 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.
growth trajectories in decoding accuracy and ORF measure is widely used to measure students' reading fluency. As mentioned previously, in the United States, the ORF measure is 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 factor 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>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td>Fall</td>
<td>Winter</td>
</tr>
<tr>
<td>Grade 2</td>
<td>Grade 2</td>
<td>1–3</td>
</tr>
<tr>
<td>ORF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTRP</td>
<td>Assessment</td>
<td></td>
</tr>
<tr>
<td>Grade 3</td>
<td>Grade 3</td>
<td>10–12</td>
</tr>
<tr>
<td>ORF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTRP</td>
<td>Assessment</td>
<td></td>
</tr>
<tr>
<td>Grade 4</td>
<td>Grade 4</td>
<td>19–21</td>
</tr>
<tr>
<td>ORF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTRP</td>
<td>Test</td>
<td></td>
</tr>
<tr>
<td>Grade 5</td>
<td>Grade 5</td>
<td>28–30</td>
</tr>
<tr>
<td>ORF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTRP</td>
<td>Test</td>
<td></td>
</tr>
</tbody>
</table>

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 as 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 tests 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

<table>
<thead>
<tr>
<th></th>
<th>Fall (reliability = .97)</th>
<th></th>
<th>Winter (reliability = .96)</th>
<th></th>
<th>Spring (reliability = .96)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RP 1</td>
<td>M</td>
<td>37.78</td>
<td>SD</td>
<td>27.96</td>
<td>[min, max]</td>
<td>[0, 162]</td>
</tr>
<tr>
<td></td>
<td>[37, 63]</td>
<td></td>
<td>27.12</td>
<td>24.37</td>
<td>[2, 162]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>1.00</td>
<td>Kurtosis</td>
<td>3.91</td>
<td>[3, 141]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.15</td>
<td></td>
<td>4.28</td>
<td>0.63</td>
<td>[4, 167]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.13</td>
<td></td>
<td>4.22</td>
<td>0.74</td>
<td>[3, 162]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.79</td>
<td></td>
<td>3.70</td>
<td>0.91</td>
<td>[8, 166]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.91</td>
<td></td>
<td>3.37</td>
<td>0.79</td>
<td>[4, 190]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.77</td>
<td></td>
<td>3.17</td>
<td>0.59</td>
<td>[3, 187]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.77</td>
<td></td>
<td>3.05</td>
<td>0.77</td>
<td>[3, 201]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>411</td>
<td>466</td>
<td>69.18</td>
<td>63.88</td>
<td>62.21</td>
</tr>
<tr>
<td></td>
<td>373</td>
<td></td>
<td>462</td>
<td>31.39</td>
<td>32.20</td>
<td>31.50</td>
</tr>
<tr>
<td></td>
<td>372</td>
<td></td>
<td>461</td>
<td>31.29</td>
<td>32.20</td>
<td>31.50</td>
</tr>
</tbody>
</table>

### Oral reading fluency: Grade 3

<table>
<thead>
<tr>
<th></th>
<th>Fall (reliability = .96)</th>
<th></th>
<th>Winter (reliability = .95)</th>
<th></th>
<th>Spring (reliability = .94)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RP 10</td>
<td>M</td>
<td>66.72</td>
<td>SD</td>
<td>33.78</td>
<td>[min, max]</td>
<td>[0, 184]</td>
</tr>
<tr>
<td></td>
<td>73.82</td>
<td></td>
<td>37.53</td>
<td>36.10</td>
<td>[0, 198]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>69.64</td>
<td></td>
<td>36.10</td>
<td>36.10</td>
<td>[0, 196]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>0.65</td>
<td>Kurtosis</td>
<td>3.25</td>
<td>[0, 190]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.52</td>
<td></td>
<td>1.50</td>
<td>2.86</td>
<td>[10, 198]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.63</td>
<td></td>
<td>3.00</td>
<td>0.26</td>
<td>[13, 220]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.26</td>
<td></td>
<td>2.93</td>
<td>0.15</td>
<td>[13, 220]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.41</td>
<td></td>
<td>2.93</td>
<td>0.41</td>
<td>[10, 229]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.41</td>
<td></td>
<td>3.42</td>
<td>0.54</td>
<td>[11, 230]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>435</td>
<td>472</td>
<td>96.20</td>
<td>88.89</td>
<td>95.46</td>
</tr>
<tr>
<td></td>
<td>434</td>
<td></td>
<td>472</td>
<td>35.10</td>
<td>33.99</td>
<td>34.98</td>
</tr>
<tr>
<td></td>
<td>432</td>
<td></td>
<td>472</td>
<td>31.50</td>
<td>33.99</td>
<td>34.98</td>
</tr>
</tbody>
</table>

### Oral reading fluency: Grade 4

<table>
<thead>
<tr>
<th></th>
<th>Fall (reliability = .93)</th>
<th></th>
<th>Winter (reliability = .93)</th>
<th></th>
<th>Spring (reliability = .95)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RP 19</td>
<td>M</td>
<td>98.70</td>
<td>SD</td>
<td>34.26</td>
<td>[min, max]</td>
<td>[4, 194]</td>
</tr>
<tr>
<td></td>
<td>109.61</td>
<td></td>
<td>37.32</td>
<td>36.59</td>
<td>[6, 214]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>95.59</td>
<td></td>
<td>36.59</td>
<td>196</td>
<td>[5, 203]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>-0.20</td>
<td>Kurtosis</td>
<td>2.77</td>
<td>[16, 204]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.11</td>
<td></td>
<td>2.76</td>
<td>-0.12</td>
<td>[17, 213]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.12</td>
<td></td>
<td>2.84</td>
<td>0.21</td>
<td>[20, 199]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.21</td>
<td></td>
<td>2.90</td>
<td>-0.14</td>
<td>[22, 221]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.08</td>
<td></td>
<td>2.57</td>
<td>0.20</td>
<td>[28, 232]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.16</td>
<td></td>
<td>2.57</td>
<td>-0.13</td>
<td>[25, 208]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>475</td>
<td>532</td>
<td>123.62</td>
<td>128.59</td>
<td>112.90</td>
</tr>
<tr>
<td></td>
<td>475</td>
<td></td>
<td>533</td>
<td>37.61</td>
<td>35.71</td>
<td>37.88</td>
</tr>
<tr>
<td></td>
<td>475</td>
<td></td>
<td>532</td>
<td>37.61</td>
<td>35.71</td>
<td>37.88</td>
</tr>
</tbody>
</table>

### Oral reading fluency: Grade 5

<table>
<thead>
<tr>
<th></th>
<th>Fall (reliability = .92)</th>
<th></th>
<th>Winter (reliability = .93)</th>
<th></th>
<th>Spring (reliability = .94)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RP 28</td>
<td>M</td>
<td>107.95</td>
<td>SD</td>
<td>29.30</td>
<td>[min, max]</td>
<td>[19, 193]</td>
</tr>
<tr>
<td></td>
<td>103.52</td>
<td></td>
<td>33.36</td>
<td>37.04</td>
<td>[20, 184]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>120.00</td>
<td></td>
<td>37.04</td>
<td>213</td>
<td>[25, 213]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>0.06</td>
<td>Kurtosis</td>
<td>3.06</td>
<td>[30, 205]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.10</td>
<td></td>
<td>2.41</td>
<td>-0.10</td>
<td>[24, 214]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.10</td>
<td></td>
<td>2.53</td>
<td>-0.99</td>
<td>[32, 224]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.09</td>
<td></td>
<td>3.00</td>
<td>-0.17</td>
<td>[26, 27]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.06</td>
<td></td>
<td>2.61</td>
<td>-0.25</td>
<td>[28, 227]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.02</td>
<td></td>
<td>2.61</td>
<td>-0.03</td>
<td>[15, 226]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>461</td>
<td>482</td>
<td>122.59</td>
<td>118.37</td>
<td>117.79</td>
</tr>
<tr>
<td></td>
<td>461</td>
<td></td>
<td>482</td>
<td>31.25</td>
<td>32.97</td>
<td>36.67</td>
</tr>
<tr>
<td></td>
<td>461</td>
<td></td>
<td>482</td>
<td>31.25</td>
<td>32.97</td>
<td>36.67</td>
</tr>
</tbody>
</table>

**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.
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 time point, 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 $\mu_B$ of the random intercept factor represents the average ORF starting level. To allow for a potential variation...
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 $\mu_{B_1}$ of the random slope factor. Variances of the residual time-specific ORF factors are constrained to be equal (i.e., parameter $\sigma^2_{B_2}$).

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 partial invariance 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: \( \chi^2(33) = 81.17, p < .001, \text{CFI} = 0.991, \text{RMSEA} = 0.053, p = .335, \text{SRMR} = 0.022; \) grade 3: \( \chi^2(36) = 118.53, p < .001, \text{CFI} = 0.984, \text{RMSEA} = 0.065, p = .018, \text{SRMR} = 0.041; \) grade 4: \( \chi^2(32) = 157.28, p < .001, \text{CFI} = 0.977, \text{RMSEA} = 0.082, p = .082, \text{SRMR} = 0.053; \) and grade 5: \( \chi^2(31) = 314.00, p < .001, \text{CFI} = 0.959, \text{RMSEA} = 0.094, p = .128, \text{SRMR} = 0.094. \)

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 \( \mu_{B_0} = 38.29 \pm 1.27, p < .001), \) which rapidly increased (i.e., random slope \( \mu_{B_1} = 26.11 \pm 0.70, p < .001 \)) across the year up to about 65 WCPM in the spring. The growth trajectory is approximately linear, with 57% (i.e., loading \( b = 0.57 \pm 0.02 \)) of the total average change in ORF in grade 2 already occurring by winter. A similar pattern of results occurred in grade 3 (\( \mu_{B_0} = 66.42 \pm 1.52, p < .001; \mu_{B_1} = 31.48 \pm 0.77, p < .001; b = 0.58 \pm 0.02 \)). In grade 4, the growth trajectory starts at about the same level (\( \mu_{B_0} = 97.39 \pm 1.44, p < .001 \)) as the spring ORF results for grade 3 but still shows continuing ORF growth (\( \mu_{B_1} = 25.92 \pm 0.76, p < .001 \)), although initially there is now a slower increase between fall and winter (\( b = 0.15 \pm 0.03 \)), with an increase to spring accounting for 85% of the average total growth. In grade 5, the ORF growth trajectory seems to decrease (\( \mu_{B_1} = 14.13 \pm 0.84, p < .001 \)), with the initial average level in the fall for grade 5 (\( \mu_{B_0} = 104.36 \pm 1.42, p < .001 \)) 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 (\( b = 1.74 \pm 0.08 \)) and not in the spring as would be expected.
### TABLE 3
Descriptive Statistics for National Tests of Reading Proficiency (NTRP) Across Grades 2–5

| Timepoint | 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 |
|           | **M** | | | | | | | | | | | | | | |
| [min, max]| [24, 25] | [5, 21] | [4, 21] | [2, 18] | [0, 10] | [3, 65] | [0, 14] | [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** | | | | **M** | | | | **M** | | | | **M** | | | |
|-----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
|           | 21.5   | 18.50  | 6.70   | 6.20   | 0.86   | 0.86   | 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.*
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 ($\sigma^2_{B_0} = 706.39 \pm 59.36$, $1,145.34 \pm 75.90$, $1,088.59 \pm 64.58$, and $967.97 \pm 61.65$, respectively) vary much more than growth rates ($\sigma^2_{B_1} = 31.86 \pm 31.30$, $65.50 \pm 29.61$, $59.59 \pm 19.24$, and $1.45 \pm 4.77$, 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 ($r_{B_0,B_1} = -.16, p = .177; r_{B_0,B_1} = .289, p = .001$). 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.
ORF factors across measurements within a grade (i.e., $r_{F_t-1} = .92–.94$, $.94–.96$, $.95–.97$, 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 assess 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 sub-skills 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 assess both predictive relations with the ORF measure (winter and fall measurement occasion).
### TABLE 5
Correlations Between All Observed Measures With the Sample Size at All Timepoints for Grade 2

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

**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

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

**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

<table>
<thead>
<tr>
<th>Measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ORF 1</td>
<td>476</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. ORF 2</td>
<td>.94</td>
<td>.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. ORF 3</td>
<td>.93</td>
<td>.94</td>
<td>.475</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. ORF 4</td>
<td>.89</td>
<td>.90</td>
<td>.91</td>
<td>532</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. ORF 5</td>
<td>.92</td>
<td>.92</td>
<td>.92</td>
<td>.93</td>
<td>533</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. ORF 6</td>
<td>.88</td>
<td>.89</td>
<td>.90</td>
<td>.92</td>
<td>.93</td>
<td>532</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. ORF 7</td>
<td>.90</td>
<td>.89</td>
<td>.89</td>
<td>.91</td>
<td>.92</td>
<td>.90</td>
<td>443</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. ORF 8</td>
<td>.90</td>
<td>.89</td>
<td>.89</td>
<td>.90</td>
<td>.91</td>
<td>.90</td>
<td>.95</td>
<td>443</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. ORF 9</td>
<td>.88</td>
<td>.88</td>
<td>.89</td>
<td>.91</td>
<td>.90</td>
<td>.91</td>
<td>.94</td>
<td>.94</td>
<td>441</td>
<td></td>
</tr>
<tr>
<td>10. NTRP</td>
<td>.52</td>
<td>.48</td>
<td>.50</td>
<td>.50</td>
<td>.48</td>
<td>.46</td>
<td>.48</td>
<td>.47</td>
<td>.50</td>
<td>165</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ORF 1</td>
<td>462</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. ORF 2</td>
<td>.92</td>
<td>.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. ORF 3</td>
<td>.93</td>
<td>.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. ORF 4</td>
<td>.89</td>
<td>.88</td>
<td>.89</td>
<td>482</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. ORF 5</td>
<td>.89</td>
<td>.88</td>
<td>.89</td>
<td>.93</td>
<td>482</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. ORF 6</td>
<td>.88</td>
<td>.90</td>
<td>.88</td>
<td>.92</td>
<td>.94</td>
<td>482</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. ORF 7</td>
<td>.88</td>
<td>.87</td>
<td>.89</td>
<td>.91</td>
<td>.91</td>
<td>.90</td>
<td>443</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. ORF 8</td>
<td>.87</td>
<td>.87</td>
<td>.88</td>
<td>.91</td>
<td>.91</td>
<td>.91</td>
<td>.95</td>
<td>443</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. ORF 9</td>
<td>.89</td>
<td>.90</td>
<td>.90</td>
<td>.92</td>
<td>.92</td>
<td>.92</td>
<td>.94</td>
<td>.94</td>
<td>443</td>
<td></td>
</tr>
<tr>
<td>10. NTRP</td>
<td>.55</td>
<td>.54</td>
<td>.55</td>
<td>.49</td>
<td>.53</td>
<td>.53</td>
<td>.49</td>
<td>.51</td>
<td>.54</td>
<td>302</td>
</tr>
</tbody>
</table>

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 \( r = .73, .75, \text{ and }.75 \text{ respectively} \) were stronger than with the three other subtests that targeted higher level reading skills (reading narrative text: \( r = .44, .47, \text{ and }.47 \text{ respectively} \) and vocabulary (word knowledge: \( r = .31, .36, \text{ and }.35 \text{ respectively} \)).

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., Garcia & 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., Silbergliit & 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

**APPENDIX A**

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

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

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

<table>
<thead>
<tr>
<th>Reading passage</th>
<th>ΔDIF*</th>
<th>Intercept</th>
<th>Loading</th>
<th>LIX readability formula</th>
<th>Flesch readability formula</th>
<th>Summary</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>0.00</td>
<td>1.02***</td>
<td>22</td>
<td>23</td>
<td>81.9</td>
<td>3.9</td>
<td>5</td>
</tr>
<tr>
<td>29</td>
<td>0.00</td>
<td>1.00</td>
<td>24</td>
<td>29</td>
<td>77.0</td>
<td>5.3</td>
<td>6</td>
</tr>
<tr>
<td>30</td>
<td>15.36***</td>
<td>1.00</td>
<td>22</td>
<td>27</td>
<td>78.3</td>
<td>5.7</td>
<td>6</td>
</tr>
<tr>
<td>31</td>
<td>-10.63***</td>
<td>1.00</td>
<td>26</td>
<td>26</td>
<td>75.2</td>
<td>5.8</td>
<td>5</td>
</tr>
<tr>
<td>32</td>
<td>-17.28***</td>
<td>1.12***</td>
<td>25</td>
<td>22</td>
<td>77.4</td>
<td>5.0</td>
<td>6</td>
</tr>
<tr>
<td>33</td>
<td>0.00</td>
<td>1.00</td>
<td>27</td>
<td>23</td>
<td>77.2</td>
<td>5.1</td>
<td>5</td>
</tr>
<tr>
<td>34</td>
<td>4.87***</td>
<td>1.00</td>
<td>19</td>
<td>23</td>
<td>76.9</td>
<td>5.0</td>
<td>5</td>
</tr>
<tr>
<td>35</td>
<td>0.00</td>
<td>1.00</td>
<td>25</td>
<td>25</td>
<td>77.0</td>
<td>5.5</td>
<td>6</td>
</tr>
<tr>
<td>36</td>
<td>0.00</td>
<td>1.00</td>
<td>26</td>
<td>30</td>
<td>73.6</td>
<td>5.9</td>
<td>6</td>
</tr>
</tbody>
</table>

*Wald tests were run for the ΔDIF parameters.

*p < .05. ***p < .001.

### APPENDIX B

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Number of items</th>
<th>Timed</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>National reading assessment subtests for grade 2: Spring 2013</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recognizing letters</td>
<td>25</td>
<td>1 minute</td>
<td>The students were presented printed capital letters and asked to identify the same printed lowercase letters and vice versa.</td>
</tr>
<tr>
<td>Writing words</td>
<td>16</td>
<td>—</td>
<td>To measure skills in spelling, the examiner asked the students to write words when listening to the teacher read words aloud.</td>
</tr>
<tr>
<td>Reading words</td>
<td>21</td>
<td>2 minutes</td>
<td>The students were asked to look at a picture and identify which of four words represented the picture.</td>
</tr>
<tr>
<td>Splitting compound words</td>
<td>21</td>
<td>1 minute</td>
<td>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.</td>
</tr>
<tr>
<td>Reading sentences</td>
<td>18</td>
<td>2 minutes</td>
<td>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.</td>
</tr>
<tr>
<td>Following written instructions</td>
<td>10</td>
<td>2.5 minutes</td>
<td>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”).</td>
</tr>
<tr>
<td>Reading text</td>
<td>6</td>
<td>—</td>
<td>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.”</td>
</tr>
</tbody>
</table>

(continued)
### National reading assessment subtests for grade 3: Spring 2013

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Number of items</th>
<th>Timed</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chains of words</td>
<td>66</td>
<td>5 minutes</td>
<td>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”).</td>
</tr>
<tr>
<td>Reading narrative text</td>
<td>9</td>
<td>15 minutes</td>
<td>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.</td>
</tr>
<tr>
<td>Word knowledge</td>
<td>20</td>
<td>—</td>
<td>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.</td>
</tr>
<tr>
<td>Reading expository text</td>
<td>7</td>
<td>15 minutes</td>
<td>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.</td>
</tr>
</tbody>
</table>