Assessing reading and spelling abilities from three different angles – correlations between test scores, teachers’ assessment and children’s self-assessments in L1 and L2 children

Lisa Paleczek*, Susanne Seifert, Susanne Schwab, and Barbara Gasteiger-Klicpera

University of Graz, Merangasse 70/2, 8010 Graz

Abstract

Aside from by standardized tests, students’ reading (decoding, reading comprehension) and spelling abilities can also be assessed by teachers or by self-assessment. This study investigated the correlations of these three types of assessment, taking children’s L1 into account. Analyses revealed that, for L1 (n=317) and L2 (n=247) children, teachers’ assessments moderately correlate with test results and do not differentiate well between the abilities. Children’s self-assessments showed lower correlations. Additionally, comparing L1 to L2 children, lower correlations were shown when L2 children assessed their decoding. Further analyses revealed that L2 children tend to overestimate their abilities. Results were discussed in the light of the importance of correct and distinct teacher judgements to provide individual support.

Keywords: accuracy of teachers’ assessments, accuracy of children’s self-assessment, reading, spelling, L1 and L2 learners, elementary school

1. Introduction

1.1 Reading and spelling abilities

The most important skills to be acquired and practiced at elementary school are the children’s reading and spelling abilities. These abilities are crucial for being able to come upon and process information successfully (David, 2010). This skill permeates the whole school career as well as the further work career (Duzy, Ehm, Souvignier, Schneider, & Gold, 2014) and therefore also plays a central role in personal well-being (EU High Level...

Reading is quite a complex ability, divided into several sub-abilities. Decoding is one of the first abilities to be acquired and thus represents the *conditio sine qua non* of the reading development process (Klicpera, Schabmann, & Gasteiger-Klicpera, 2010). This ability can be subdivided into non-lexical and lexical decoding. As lexical decoding progresses, the number of words stored in the so-called mental lexicon increases. The fast and increasingly automatic retrieval of words from this lexicon eases and fastens the reading process and is required for efficient reading comprehension (Klicpera et al., 2010). The better developed this decoding process is, the more processing resources are available for reading comprehension (Perfetti, 1985). Decoding has been shown to be a strong predictor for word-level reading comprehension (Wagner et al., 1997).

Reading and spelling are two highly connected abilities as well. In her model of reading and spelling development, Frith (1985) assumes that spelling skills do not develop independently from reading skills. As shown before, especially in the early grades, spelling abilities are highly inter-correlated with reading skills (Bishop & Leonard, 2000; Cunningham & Stanovich, 1993).

In Austria, 20% of ten-year-old students do not even possess basic competences in reading, whereas just 5% show high performance in this area (e.g. Bergmüller & Wintersteller, 2012; Herzog-Punzenberger, Bruneforth, & Lassnig, 2013). Poor reading abilities in children lead to a high risk regarding the successful completion of their school career (Lonigan, Purpura, Wilson, Walker, & Clancy-Menchetti, 2012). Children with a different first language from the language of instruction (L2 children) show a higher risk for developing problems at school, as language is the foundation of most of the subjects (Herzog-Punzenberger & Schnell, 2013). Especially regarding reading acquisition, L2 children show more deficits than their L1 peers (Roberts, Sorajani, & Vaughn, 2010; Schabmann, Landerl, Bruneforth, & Schmidt, 2013). This could also be shown for L2 children in Austria. Approximately 30% of the differences in achievement between L1 and L2 children can be explained by differences in their family’s cultural capital (Bergmüller & Herzog-Punzenberger, 2012).

Results like the ones mentioned are usually retrieved from international student assessments (e.g. PISA: Schwantner, Toferer, & Schreiner, 2012; or PIRLS: Suchań, Wallner-Paschon, Bergmüller, & Schreiner, 2012), where information about reading and spelling abilities is typically obtained by using standardized tests. These tests are often quite time-consuming and can only be applied by especially trained people.

1.2 Accuracy of teachers’ assessment

As the application of standardized tests is not very common in daily classroom life, teachers’ diagnostic competencies are essential (Artelt & Gräsel, 2009). Teachers need to be able to determine students’ abilities adequately in order to offer an appropriate learning environment for all children in the reading classroom (Kame‘enui et al., 2006). Moreover, teachers’ assessment of students’ abilities can influence the development of these abilities. Schwab, Seifert and Gasteiger-Klicpera (submitted) demonstrated this effect specifically for the development of reading comprehension in second graders. As could be shown already, teachers are not always able to accurately assess the abilities of their students and tend to make inaccurate judgements about their students’ reading abilities (e.g. Begeny, Krouse, Brown, & Mann, 2011). There are several biases in teachers’ judgements of student performance. The Big-Fish-Little-Pond Effect (Südkamp & Möller, 2009), anchoring effects (Dünnebier, Gräsel, & Krolak-Schwerdt, 2009) or teachers’ goals and expertise (Krolak-Schwerdt, Böhmer, & Gräsel, 2012) are just some of the possible biases. The correlations between teachers’ assessment of reading and students’ abilities measured by standardized tests can be described as moderate, although results are inconsistent (Begeny et al., 2011: $r=.43$ to 58, $p<.01$; Demaray & Elliott, 1998: $r=.82$; Karing & Artelt, 2013: $r=.34$, $p<.01$). On the one hand, research has shown that teachers assess more accurately when being asked to assess specific skills (Feinberg & Shapiro, 2009); on the other hand, more recent research suggests that global judgements of teachers have higher accuracy than task-specific judgements (Karing, Matthäi, & Artelt, 2011). Especially when reading and/or spelling abilities are impaired, teachers’ assessments of these skills have only limited validity (Herzog-Punzenberger et al., 2012; Schmidt & Schabmann, 2010).

The fact that teachers sometimes tend to misclassify children regarding their reading abilities can lead to the assumption that teachers’ judgements are occasionally influenced by certain child characteristics which do not necessarily influence the children’s actual performance. Research showed that, in addition to emergent literacy
skills, perceptions of classroom behaviour, too, had an influence on teachers’ judgements: students who showed behavioural issues were judged to show poorer academic outcomes, regardless of their actual scholastic skills (Bennett, Gottesman, Rock, & Cerullo, 1993; Hecht & Greenfield, 2002). In a study examining the sensitivity of teacher ratings of L1 and L2 children for reading disabilities, Limbos and Geva (2001) found that only in the long term the accuracy of teachers’ assessments was comparable for the two groups. However, when measuring the first time, sensitivity of teacher ratings were higher for L1 children. Hence, there was a tendency to misclassify L2 children as having a reading disability due to an overreliance on oral language performance.

1.3 Children’s self-assessment

Abilities in reading can also be assessed from a different angle, namely through children’s self-assessment. Since there is a close link between self-assessment and academic self-concept (Marsh & Craven, 2006), the children’s perspective on their abilities should be taken into account. One aspect, which could influence children’s abilities in general, is their self-concept of their own abilities. Schwab, Seifert and Gasteiger-Klicpera (submitted) showed that, especially for poor readers, the self-estimation of one’s own abilities influences the development of reading comprehension. More specifically, it could be shown that how children perceive themselves in terms of reading abilities is significantly correlated with their reading achievements. In line with this, Spinath et al. revealed that for predicting students’ achievement in mathematics and English, the children’s self-perception about their abilities played an important role (Spinath, Spinath, Harlaar, & Plomin, 2006). However, other studies revealed that if any inter-correlations between children’s self-assessment and their standardized test-scores existed, they were only quite low ($r=.29$, $p<.05$; Lynch, 2012). With regard to Austria, studies showed that nearly 50% of the children in fourth grade have a positive self-concept about their reading abilities. Test scores, on the other hand, show that this self-perception is not necessarily accurate: e.g. 10% of Austrian fourth graders with rather poor achievement in reading tests are convinced of their reading abilities (Wallner-Paschon, 2012).

Children base their self-perception on different sources. Their perception is influenced by teachers’ feedback about their performance or progress (marks, verbal feedback, etc.), by comparing themselves to their classmates (Hellmich & Günther, 2011) but also by other characteristics (e.g. gender and age: Bouffard, Marcoux, Vezeau, & Bordeleau, 2003). Moreover, training children’s self-assessment can be used in the classroom as a kind of formative tool, if reinforced by the teachers (Bingham, Holbrook, & Meyers, 2010).

As research produced contradictory results about whether or not children are able to properly describe their competencies in the scholar context, researchers started to consider certain child characteristics (age, gender). Furthermore, studies began to differentiate between the domains (subjects at school) that were investigated. Considering this, only girls (2nd grade) showed significant inter-correlations between their self-perception of competence and their marks in reading ($r=.29$, $p<.05$) and mathematics ($r=.41$, $p<.01$). Later, in 3rd grade, boys showed inter-correlations in mathematics, but not in reading (boys: mathematics: $r=.30$, $p<.05$; girls: reading: $r=.32$, $p<.05$; mathematics: $r=.46$, $p<.001$) (Bouffard, Marcoux, Vezeau, & Bordeleau, 2003).

With respect to child characteristics, no results concerning differences in L1 and L2 children could be found.

1.4 Purpose of the study

Since the specification of the investigated domain apparently matters, in the present study, the domain of German language arts is to be examined. More specifically, in the present study, the accuracy of teachers’ assessment and children’s self-assessment of reading and spelling abilities at the beginning of third grade is to be investigated. Even though various studies exist focusing on this topic, none of these studies analysed if these assessments differ in their accuracy when considering the children’s first language. Three sub-areas of abilities will be analysed, namely decoding, reading comprehension and spelling. The paper aims at addressing two groups of research questions. The first group concerns an overall view of the entire sample.

To that end, four different research questions related to the whole sample were asked:
1. How well do the different types of assessment (standardized tests, teachers and children) differentiate the three abilities (decoding, reading comprehension and spelling) in question?
2. How accurately do teachers assess their students in the above-mentioned abilities?
3. How accurately do children perceive themselves in the above-mentioned abilities?
4. Do teachers’ and children’s assessments about these three abilities correlate?

Further attention was given to examining whether there are differences regarding child characteristics. Thus, differences between L1 and L2 children were studied. Hence, in a second step, the same four research questions above mentioned were asked again, this time dividing the sample in L1 and L2 children (2nd group, research questions 5-8). The two groups (L1 and L2 children) were compared with regard to significant differences.

2. Methods

2.1 Sample

The sample consisted of 564 children who were in 15 different schools, distributed among 32 classrooms (therefore, 32 teachers). This sample was special in that all these classrooms had at least 30% of L2 children. The children were assessed at the beginning of the third grade (age: \( M = 8.4, SD = 0.5 \)). 52.1% \((n = 311)\) of the children were boys and 47.9% \((n = 293)\) were girls. 44.5% \((n = 269)\) were L2 children. This group consisted of 22 different languages. The biggest language group among the L2 children was Bosnian, Serbian or Croatian with 9.3% of the children \((n = 56)\). 5.3% \((n = 32)\) spoke Chechen, 4.8% \((n = 29)\) Albanian, 4.3% \((n = 26)\) Turkish and 3.1% \((n = 19)\) Arabic as a first language.

2.2 Instruments and procedures

Reading and spelling tests

Standardized tests gathered information about the children’s reading skills (divided into decoding and reading comprehension) and spelling skills. To assess decoding skills, the reading part of the Salzburg Reading and Spelling Test (SLRT II: Moll & Landerl, 2010) was deployed. The SLRT II is an individual reading test which assesses decoding of non-words and decoding of words. This way, the test allows a separate diagnosis of phonological decoding (sub-scale Non-word Decoding) and lexical reading (sub-scale Word Decoding). To assess the children’s reading comprehension, the Reading Comprehension Test for First to Sixth Graders (Elfe 1-6: Lenhard & Schneider, 2006) was applied. This test measures three different levels of reading comprehension: word, sentence and text level. To get an overall reading comprehension score, raw scores of each level were z-transformed and a composite sum of the three z-scores was formed. Further analyses were conducted with this overall reading comprehension score. Spelling abilities were tested with the Hamburg Spelling Test (HSP 1-9: May, 2002). This test counts correctly written words and graphemes. For the present study, the number of correctly written graphemes was used for further analyses because they offer more precise information.

Teachers’ assessments

Teachers assessed each child’s decoding, reading comprehension and spelling abilities on a 4-point Likert scale (1 = bad to 4 = good) (e.g. “Please assess the child’s spelling ability”).

Children’s self-assessments

Corresponding to a frequently used one-item measure in this research field, children were asked to assess their reading fluency (that highly corresponds to decoding skills) and reading comprehension as well as their spelling skills (“I read very fast.”; “I easily understand stories I am reading.”; “I am good at spelling.”). Children were rating these statements on a 3-point Likert scale, ranging from total agreement (3) to total disagreement (1).

Table 1. Means and standard deviations of the test scores achieved, teachers’ assessments and children’s self-assessment of decoding, reading comprehension and spelling (N=564).

<table>
<thead>
<tr>
<th></th>
<th>( M ) (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decoding score Non-words/ Words (items read per minute)</td>
<td>35.47 (11.15) / 52.25 (19.06)</td>
</tr>
<tr>
<td>Reading Comprehension score (z-score)</td>
<td>- .25 (2.85)</td>
</tr>
<tr>
<td>Spelling score (148 maximum possible)</td>
<td>132.6 (15.17)</td>
</tr>
<tr>
<td>Teacher-assessed Decoding (4-point Likert scale)</td>
<td>3.06 (0.91)</td>
</tr>
<tr>
<td>Teacher-assessed Reading Comprehension (4-point Likert scale)</td>
<td>3.15 (0.94)</td>
</tr>
<tr>
<td>Teacher-assessed Spelling (4-point Likert scale)</td>
<td>2.87 (0.96)</td>
</tr>
<tr>
<td>Children’s self-assessed Decoding (3-point Likert scale)</td>
<td>2.47 (0.55)</td>
</tr>
<tr>
<td>Children’s self-assessed Reading Comprehension (3-point Likert scale)</td>
<td>2.67 (0.53)</td>
</tr>
<tr>
<td>Children’s self-assessed Spelling (3-point Likert scale)</td>
<td>2.66 (0.50)</td>
</tr>
</tbody>
</table>
Procedures

First, children had to rate their own abilities. Then, two standardized tests (reading comprehension: *Elfe 1-6* and spelling: *HSP 1-9*) were administered in the classroom. The size of the group was determined by the number of children in class (9-25 children). Thereafter, the decoding test (*SLRT II*) was administered individually. Meanwhile, the respective teacher filled out questionnaires about each child’s abilities.

3. Results

3.1 Correlations within the entire sample

In order to analyse the relationships between standardized test-scores, teachers’ and children’s assessment, respectively (see Table 2), correlations were conducted. As the data were not normally distributed (only reading comprehension was), Spearman correlations were performed on the entire sample. Subsequently, correlations were compared. To find out if the correlations differed significantly from each other, Fisher’s z-transformation was used according to Fisher’s z-table (Sedlmaier & Renkewitz, 2013, p.905).

To address the first research question, the inter-correlations within the three types of assessment were analysed (see triangles in Table 2).

*First*, the *inter-correlations between the standardized test-scores* (see first triangle in the left corner, Table 2), are moderate to rather high ($r_S = .52$ to $.74$, $p < .01$). The highest inter-correlation was found between word decoding and reading comprehension ($r_S = .74$, $p < .01$). The lowest inter-correlation existed between non-word decoding and spelling, still being highly significant ($r_S = .52$, $p < .01$). The mean of these inter-correlations was calculated by using z-transformations ($\bar{z} = .74$; $\bar{r}_S = .63$).

*Second*, the *inter-correlations between teachers’ assessments* (see second triangle in the middle, Table 2) of the three different abilities (decoding, reading comprehension, spelling) are also rather high, again all being highly significant ($r_S = .62$ to $.71$, $p < .01$). The mean was calculated (Fisher’s $\bar{z} = .81$, $\bar{r}_S = .67$).

*Third*, the *inter-correlations between children’s self-assessments* (see third triangle in the right bottom corner, Table 2) were analysed. In contrast to the two previously mentioned inter-correlations (within teachers’ assessment and within standardized tests), the children’s self-assessments only showed one highly significant inter-correlation (for decoding x reading comprehension: $r_S = .24$, $p < .01$). Still, the mean was calculated ($\bar{z} = .12$; $\bar{r}_S = .12$).

To find out if the inter-correlations within the three kinds of assessment differed from each other, means of the three inter-correlations were analysed by using following formula (Bortz, 1989, p.265):

$$z = \frac{(Z_1 - Z_2)}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 2}}}$$

Analyses showed that the mean inter-correlation of teachers’ assessments ($\bar{r}_S = .67$) is not significantly higher ($\sigma(Z_1 - Z_2) = 0.05971$; $z = 1.13$) than the mean inter-correlation of standardized tests ($\bar{r}_S = .63$). The mean inter-correlation of standardized test scores and the mean of inter-correlations in teachers’ assessments are significantly higher ($z = 10.36$ and $z = 11.49$, respectively) than the mean inter-correlation of children’s assessments ($\bar{r}_S = .12$). This means that test scores and teachers’ assessment differentiate to the same extend, whereas children dissociate the three abilities to a higher extent. However, it might be that children dissociate the three abilities more than necessary.

Moreover, research questions two and three aimed at the accuracy of teachers’ and children’s assessment. Therefore their correlations with standardized test-scores were analysed (see upper squares in Table 2).

*First*, in order to answer research question 2, the correlations between *standardized test scores and teachers’ assessments* (see square on the left) were analysed. Overall (see in bold within this square), these correlations with the corresponding test scores were of medium size but all highly significant ($r_S = .50$ to $.60$, $p < .01$). This means that teachers’ accuracy is medium for all the assessed abilities. When looking separately at the three abilities assessed by teachers (see columns within this square), the following results were found.

In terms of decoding, teacher-assessed decoding x decoding test-score is not the highest correlation in the column. This shows that when teachers are asked to assess decoding, they do not distinguish this ability from reading...
comprehension and spelling. Besides, teachers’ assessment of decoding correlates significantly higher ($z=2.41$) with the word decoding score ($r_S=.60, p<.01$) than with the non-word decoding score ($r_S=.50, p<.01$) of the standardized tests. This means that the teachers assess a general factor of reading and do not assess individual abilities. This may be especially crucial for children with dyslexia since they show very different profiles and need help in specific aspects according to their difficulties. For example, they often have special difficulties in phonological recoding and in decoding non-words.

Table 2. Correlations between decoding, reading comprehension and spelling; for all three types of assessment.

<table>
<thead>
<tr>
<th></th>
<th>Standardized</th>
<th>Teachers</th>
<th>Children</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>Standardized</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) Decoding</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-word/Word</td>
<td>-.61**/.74**</td>
<td>.50**/.60**</td>
<td>.32**/.40**</td>
</tr>
<tr>
<td>(2) Reading c.</td>
<td>-</td>
<td>.63**</td>
<td>.60**</td>
</tr>
<tr>
<td>(3) Spelling</td>
<td></td>
<td>.57**</td>
<td>.42**</td>
</tr>
<tr>
<td>Teachers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) Decoding</td>
<td></td>
<td>.71**</td>
<td>.62**</td>
</tr>
<tr>
<td>(2) Reading c.</td>
<td>-</td>
<td>.28**</td>
<td>.10**</td>
</tr>
<tr>
<td>(3) Spelling</td>
<td></td>
<td>.31**</td>
<td>.14**</td>
</tr>
<tr>
<td>Children</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) Decoding</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) Reading c.</td>
<td>-</td>
<td>.24**</td>
<td>.03</td>
</tr>
<tr>
<td>(3) Spelling</td>
<td></td>
<td>.28**</td>
<td>.20**</td>
</tr>
</tbody>
</table>

N=564; Reading c. = reading comprehension; ** p<.05; * p<.01 two-tailed

With respect to the teacher-assessed reading comprehension, the correlation with the reading comprehension test score is the highest ($z=4.14$; comparing $r_S=.60$ with $r_S=.42$). This means that teachers seem to accurately assess reading comprehension and do distinguish this ability adequately from decoding and spelling. In spelling, a similar pattern as in decoding could be revealed. Even though teacher-assessed spelling x spelling test score represents the highest correlation in this column, it does not differ significantly from teacher-assessed spelling x reading comprehension test score. This shows that when teachers are asked to assess spelling, they seem to distinguish spelling from decoding, but assess reading comprehension simultaneously.

In conclusion, only teacher-assessed reading comprehension is depicting the same ability as reading comprehension assessed by standardized tests. In contrast, the other two teacher-assessed abilities (decoding and spelling) also depict abilities that are not corresponding. Therefore, it seems that teachers are not able to assess children’s decoding and spelling independently from the other abilities.

Second, to answer research question 3, we analysed the correlations between the standardized test-scores and the children’s self-assessment (see square on the right). In general (see in bold within this square), these correlations with the corresponding test scores were lower ($r_S=.20$ to .38; $p<.01$) than the correlations between the ones observed before (standardized test scores x teachers’ assessment). This means that children do not assess their abilities as accurately as teachers do. When looking separately at the three abilities assessed by children (see columns within this square), the following results were found.

With respect to decoding, one could observe a similar pattern in the children’s assessment as in the teachers’ assessment. Even though the correlation between children-assessed decoding x decoding test score represents the highest in this column, it does not differ significantly from the other correlations in this column. This shows that when children are asked to assess decoding, they seem to assess not only decoding but also simultaneously reading comprehension and spelling. Unlike the teachers’ assessment, children’s assessment of decoding correlates equally with non-word and word decoding test scores. Although the correlation between children-assessed reading comprehension x reading comprehension test scores represents the highest in this column, no significant differences could be found from the other correlations in this column. The same can be seen in children-assessed spelling which seems to correlate highest with spelling test scores, but there were no significant differences to the other correlations in this column.

In summary, the children’s assessments never only correlate with the corresponding test scores but always with the test scores of the other two abilities as well. Thus, children are not able to evaluate their own decoding, reading
comprehension and spelling abilities differentiated from and independently of the other abilities.

Research question four aimed at the correlations between teachers’ and children’s assessment (see bottom square in Table 2). In general, these correlations (see in bold within this square) are quite low ($r_S$=.19 to .31, $p<.01$). When looking separately at the three abilities assessed by children (see columns within this square), the following results were found. Only in one of the columns, a significant difference could be found. Teacher-assessed spelling x children-assessed spelling differed significantly ($z=3.05$) from teacher-assessed reading comprehension x children-assessed spelling but not from teacher-assessed decoding x children-assessed spelling. All in all, this means that teachers’ and children’s assessments partly coincide only when assessing spelling.

3.2 Comparison of L1 and L2 children

Regarding the second group of research questions (questions 5-8), namely whether there are differences in the questions between L1 and L2 children mentioned earlier, the very same correlations were performed, this time separately for L1 and L2 children (see Table 3).

In general, many of the patterns mentioned earlier could be observed again.

Firstly, looking at the inter-correlations within the different types of assessment (see triangles, Table 3), standardized tests (triangles on the left side) showed quite high inter-correlations in both groups (L1: $r_S$=.57 to .78, $p<.01$; L2: $r_S$=.50 to .71, $p<.01$). No differences in mean correlations of test scores between L1 ($\bar{r}_S$=.67) and L2 ($\bar{r}_S$=.61) could be found. This was also true for teachers’ assessments (L1: $r_S$=.56 to .65, $p<.01$; L2: $r_S$=.56 to .70, $p<.01$, triangles in the middle). No differences in the mean correlations of teachers’ assessments between L1 ($\bar{r}_S$=.65) and L2 ($\bar{r}_S$=.61) could be found. Again in contrast to these high inter-correlations within test scores and within teachers’ assessments, children’s assessments (triangles on the right side) showed mainly non-significant correlations or they were significant but rather low (L1: $r_S$=.22, $p<.01$; L2: $r_S$=.17 to .26, $p<.01$). No differences in mean correlations of children’s assessments between L1 ($\bar{r}_S$=.17) and L2 ($\bar{r}_S$=.08) could be found.

Like before, mean inter-correlations within standardized tests do not differ significantly from mean inter-correlations within teacher assessments. This is true for L1 and L2 children. However, mean inter-correlations within test scores as well as mean inter-correlations within teachers’ assessments differ significantly from mean inter-correlations within children’s assessments. This is true for L1 ($z=8.62$; $z=7.39$) and L2 ($z=6.25$; $z=6.96$).

In a second step, correlations of the L1 group (grey in Table 4) were compared with those of the L2 group (white in Table 4). Only those correlations concerning the accuracy of children’s and teachers’ assessment were of interest (bold in Table 3).

### Table 3. Correlations between decoding, reading comprehension and spelling; for all three types of assessment; separately for L1 and L2 children.

<table>
<thead>
<tr>
<th>Standardized</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decoding</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-word/word</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading c.</td>
<td>.58**/.61</td>
<td>.61*</td>
<td>.51**</td>
<td>.64**</td>
<td>.59*</td>
<td>.53**</td>
<td>.53**</td>
<td>.59*</td>
<td>.53**</td>
</tr>
<tr>
<td>Spelling</td>
<td>.50**/.61</td>
<td>.61**</td>
<td>.51**</td>
<td>.64**</td>
<td>.59*</td>
<td>.53**</td>
<td>.53**</td>
<td>.59*</td>
<td>.53**</td>
</tr>
<tr>
<td>Teachers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decoding</td>
<td>.49**/.55**</td>
<td>.58*</td>
<td>.53**</td>
<td>.58*</td>
<td>.53**</td>
<td>.58*</td>
<td>.53**</td>
<td>.58*</td>
<td>.53**</td>
</tr>
<tr>
<td>Reading c.</td>
<td>.32**/.39**</td>
<td>.54**</td>
<td>.41**</td>
<td>.54**</td>
<td>.41**</td>
<td>.54**</td>
<td>.41**</td>
<td>.54**</td>
<td>.41**</td>
</tr>
<tr>
<td>Spelling</td>
<td>.40**/.46**</td>
<td>.53**</td>
<td>.56**</td>
<td>.53**</td>
<td>.56**</td>
<td>.53**</td>
<td>.56**</td>
<td>.53**</td>
<td>.56**</td>
</tr>
<tr>
<td>Children</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decoding</td>
<td>.26**/.31**</td>
<td>.19**</td>
<td>.27**</td>
<td>.24**</td>
<td>.16**</td>
<td>.16**</td>
<td>.24**</td>
<td>.16**</td>
<td>.16**</td>
</tr>
<tr>
<td>Reading c.</td>
<td>.05/.11</td>
<td>.19**</td>
<td>.09</td>
<td>.18**</td>
<td>.19**</td>
<td>.19**</td>
<td>.18**</td>
<td>.19**</td>
<td>.19**</td>
</tr>
<tr>
<td>Spelling</td>
<td>.01/.04</td>
<td>.00</td>
<td>.01</td>
<td>.13**</td>
<td>.06</td>
<td>.22**</td>
<td>.13**</td>
<td>.06</td>
<td>.22**</td>
</tr>
</tbody>
</table>

Reading c. = reading comprehension; grey= L1 children ($n=317$); white=L2 children ($n=247$); * $p<.05$ ** $p<.01$ two-tailed; correlations for further analyses bold

As can be seen in Table 4, only one significant difference could be found. Only the comparison of L1 and L2 children regarding the correlation between their test scores and self-assessments of decoding differed significantly. This was only true for the subscale Non-word Decoding but not for the subscale Word Decoding, although a trend
was also seen in the latter.

Table 4. Differences between observed correlations; separately for L1 and L2 children.

<table>
<thead>
<tr>
<th></th>
<th>L1 (n=317)</th>
<th>L2 (n=247)</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standardized tests x teachers’ assessment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decoding Non-word/Word</td>
<td>.53/.57</td>
<td>.49/.55</td>
<td>.59/.35</td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>.53 .59</td>
<td>.54 .60</td>
<td>.12</td>
</tr>
<tr>
<td>Spelling</td>
<td>.61 .71</td>
<td>.56 .63</td>
<td>.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Standardized tests x children’s assessment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decoding Non-word/Word</td>
<td>.41/.44</td>
<td>.26/.31</td>
<td>1.99*1.76</td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>.15 .15</td>
<td>.19 .19</td>
<td>.047</td>
</tr>
<tr>
<td>Spelling</td>
<td>.03 .03</td>
<td>.01 .01</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Teachers’ x children’s assessment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decoding</td>
<td>.37 .39</td>
<td>.24 .25</td>
<td>1.64</td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>.16 .16</td>
<td>.19 .19</td>
<td>.035</td>
</tr>
<tr>
<td>Spelling</td>
<td>.19 .19</td>
<td>.22 .22</td>
<td>.35</td>
</tr>
</tbody>
</table>

For the calculations $\sigma(Z_1-Z_2)=0.0853$ was used; grey= L1 children (n=317); white=L2 children (n=247); * $z>1.96$

To get further information about the differences, variance analyses were conducted for decoding, more precisely decoding test scores (Non-word and Word) and self-assessment of decoding. There was a significant difference between L1 and L2 children concerning their test scores in decoding ($F_{562,2}=9.63$, $p<.01$). Scheffé-tests revealed that this difference was only significant for the subscale Word Decoding and not for Non-word Decoding. L1 children showed higher test scores ($M=54.43$; $SD=19.79$) than L2 children ($M=49.45$; $SD=17.71$).

4. Discussion

This study aimed at investigating the accuracy of teachers’ assessment and children’s self-assessment of the children’s reading (decoding and reading comprehension, respectively) and spelling abilities. In addition, comparisons between L1 and L2 children were made to find out whether the correlations differ for L1 and L2 children.

*Firstly*, it was shown that these three abilities are highly inter-correlated. As expected because of the high predictive value of decoding for reading comprehension (e.g. Wagner et al., 1997) within standardized test scores, a rather high inter-correlation ($r_S=.74$, $p<.01$) between the subscales word decoding and reading comprehension could be shown. In addition, it could be shown that teachers’ assessments of the three abilities (decoding, reading comprehension and spelling) inter-correlate as highly with each other as the standardized test scores do. This result indicates that teachers, to the same degree as standardized tests, do not completely distinguish the abilities. Concerning decoding and reading comprehension, high inter-correlations could be shown already. However, it was quite surprising that neither standardized tests nor teachers assess spelling independently from reading abilities. Perhaps one can argue that it could be proven already that reading and spelling abilities are closely related to each other. This is especially true for the first grades (Cunningham & Stanovich, 1993). The analysed sample was tested at the beginning of third grade, where obviously such inter-correlations still exist.

While teachers’ as well as standardized assessment of the three abilities highly inter-correlate within the type of assessment, the children’s self-assessed reading and spelling abilities only show rather low inter-correlations. This result suggests that children seem to distinguish more strongly between the three abilities. Comparing to the test scores and teachers’ assessments, children might even perceive the three abilities in question as being too independent from each other.

*Secondly*, teachers’ accuracy was analysed by relating teachers’ assessment to test scores. It was shown that teachers do not always differentiate accurately between abilities. Teachers only showed accurate assessment when assessing reading comprehension. With respect to decoding and spelling, teachers did not show a differentiated view. In combination with the before-mentioned result, this finding suggests once again that differentiation by teachers is not based on a clear understanding of the students’ learning process. Thus, teachers tend to assess some kind of general achievement in the subject German language arts. This seems to be enough if arguing that teachers just need to grade their children in the subject German language arts in a general way. Nevertheless, to properly
support children in their reading and spelling acquisition, it is important to distinguish between these three abilities. The differentiation between abilities in decoding, reading comprehension and spelling is important to understand causes of failure in reading or spelling performance. To successfully identify children with potential problems, accurate teachers’ assessments are crucial for their diagnostic competence of reading difficulties. Children with difficulties especially rely on a teaching approach that is well-structured in manageable steps (Schabmann et al., 2012). Adaptive and therefore effective instruction in the reading classroom influences the children’s achievement and success at school (Karing & Artelt, 2013).

Another finding was that teacher’s assessment of decoding correlates more highly with the subscale Word Decoding than with the Non-word Decoding test score. This result indicates that teachers focus more on lexical decoding than on phonological one. Yet, as phonological decoding is an important step in reading acquisition and the foundation of word reading and therefore reading comprehension processes (Perfetti, 1985), it is quite important to consider phonological decoding, too.

Looking at the rather low correlations between test scores and children’s self-assessment, it seems that children at this age have no clear idea about their different abilities. This finding is underlined when looking at the low correlations between children’s self-assessment and teachers’ assessment. The highest correlation could be found between word decoding test scores and children-assessed decoding. This shows that children also seem to focus more on lexical decoding rather than on phonological decoding. Self-assessed spelling showed no significant correlation to spelling test scores. This result leads to the assumption that children have quite big difficulties in judging their spelling abilities. However, children’s assessment of spelling correlates significantly with teachers’ assessment of the same ability. This means that children and teachers agree more about the children’s spelling abilities than children agree with test scores. This shows that children’s self-assessments seem to be influenced by the teachers’ opinions. Bearing in mind that research already revealed a tendency towards misclassifications by teachers (Herzog-Punzenberger et al., 2012; Kame’enui et al., 2006; Schmidt & Schabmann, 2010), relying on teachers’ judgements might entail the risk that children do not see abilities they actually have. However, the children’s self-concept relates to their development concerning their abilities (Spinath et al., 2006).

With regard to the second group of research questions, namely if there are differences between L1 and L2 children, similar patterns could be found in general. First of all, teachers assess children’s abilities in decoding, reading and spelling independently of the children’s first language. This child characteristic does not seem to influence teachers’ accuracy of judgement.

Still, one difference could be found concerning the children’s self-assessment: L2 children show significantly lower correlations than L1 children between their self-assessed decoding and the test scores of the same ability. Further analyses revealed that L2 children tend to overestimate their abilities. In the subscale of word decoding, L2 children showed poorer achievements than their L1 peers. Yet, when looking at their self-assessment, they showed no significant differences. This means that L2 children perceive themselves as equally good at decoding as their L1 peers. However, test scores show a different picture: L2 children are weaker word decoders than L1 children. A similar phenomenon has been shown in adult L2 learners. The more advanced they are, the more accurately they judge their skills (Brantmeier, Vanderplank, & Strube, 2011).

It should be mentioned, however, that the present study has some limitations. The observed inaccuracy of teachers’ assessment might be due to question characteristics concerning how teachers’ judgement was asked for. The questions about the children’s achievement were asked in a quite general way. Some research showed accuracy rising when asking teachers in a more specific way (Feinberg & Shapiro, 2009). In contrast, Karing and Artelt (2013) revealed that teachers showed more accuracy when assessing general abilities rather than specific ones. However, in further studies, more specific questions could be asked to see if the present results remain and if there are still no differences between L1 and L2 children.

Children’s assessment only showed little accuracy. This could also be due to question characteristics. On the one hand, in order to ease the assessment for children, only a 3-point Likert scale was used. Maybe this was not enough to receive differentiated self-assessments. On the other hand, it might be questioned whether children at this age are able to understand the difference between the three abilities.

Another limitation that needs to be mentioned is that L1 and L2 learners were not comparable in some of their abilities. The higher scores of L1 children in the subscale Word Decoding could also be an artefact of different
ability levels instead of being due to a different first language. This raises the question why the correlations between the three different assessments sometimes vary between L1 and L2. Maybe other characteristics that are relevant for scholastic abilities also exert an influence. Aspects like cultural capital, socioeconomic background or IQ should therefore be controlled in further research. In addition, L1 children in the present sample often belonged to a rather low socioeconomic background and might not represent the typical L1 children.

Further research should be conducted concerning L1 and L2 children’s self-assessment in the area of reading. The different accuracy of children’s self-assessment leads to the question of where this different self-perception stems from as well as where it might lead, and these questions needs to be examined in further studies.

Acknowledgements
The authors gratefully acknowledge the cooperation and support of the federal state of Styria (Land Steiermark).

References
Feinberg, A.B., & Shapiro, E.S. (2009). Teacher Accuracy: An Examination of Teacher-Based Judgments of Students’ Reading With Differing Achievement Levels. Journal of Education Research, 102 (6), 453-462.


