What Methods of Scoring Young Children’s Spelling Best Predict Later Spelling Performance?

Treiman, Rebecca; Kessler, Brett; Caravolas, Marketa

Journal of Research in Reading

DOI:
10.1111/1467-9817.12241

Published: 01/02/2019

What Methods of Scoring Young Children’s Spelling Best Predict Later Spelling Performance?

Abstract

Background

Children's spellings are often scored as correct or incorrect, but other measures may be better predictors of later spelling performance.

Method

We examined seven measures of spelling in Reception Year and Year 1 (5–6 years old) as predictors of performance on a standardized spelling test in Year 2 (age 7).

Results

Correctness was the best predictor of later spelling by the middle of Year 1, and it significantly outperformed a binary measure of phonological plausibility at the end of Reception Year. Nonbinary measures based on Levenshtein distance were significant predictors of later spelling in the middle of Reception Year and in children who produced no correct spellings. Some widely used scales performed less well with children who did not yet produce any correct spellings.

Conclusions

Nonbinary measures of spelling performance can predict later spelling performance, but for a more restricted period than anticipated based on many theories.

Keywords: spelling; phonology; orthography; spelling errors; Levenshtein

Highlights

What is already known about this topic

- Spelling is an important skill
- Some researchers have suggested that correctness is not a good measure of spelling performance for children in early primary grades

What this paper adds

- By age 6, correctness is a good predictor of later spelling performance
- Before this point, nonbinary measures based on Levenshtein distance are significant predictors

Implications for theory, policy or practice

- Knowledge of spelling conventions emerges early in the course of spelling development
- Children’s knowledge of these conventions can be captured by correctness or computer-scored nonbinary measures
- Early assessment of spelling can help to determine which children may go on to develop difficulties

What Methods of Scoring Young Children’s Spelling Best Predict Later Spelling Performance?

The ability to spell individual words is an important foundation for writing and reading (e.g., Graham, Harris, & Chorzempa, 2002). How can we tell whether a young child is on the path toward becoming a good speller? The standardized spelling tests that are often used for educational and research purposes score responses as correct or incorrect. However, many researchers have suggested that correctness is not a good measure of performance for children in the early primary grades (e.g., Landerl & Wimmer, 2008; Ritchey, Coker, & McCraw, 2010). Theories of spelling and reading development according to which young children rely predominantly on phonological strategies (Ehri, 1986, 2005, 2014; Frith, 1985) suggest that it is not until around Grade 2 (around 8 years old in the U.S.) that children store in memory information about which sound-to-letter correspondences are appropriate for specific words or contexts within words. According to these views, measures that accept letters such as ‹k› in *camp*, on the grounds that this letter represents the phoneme /k/ in some English words, would be good indicators of a child’s current knowledge and future performance. They should perform better than measures that require orthographic correctness, downgrading such spellings as ‹kamp›. The goal of the current study was to test this idea.

Several investigators have developed methods of scoring children’s spelling productions as alternatives to the traditional correct versus incorrect measure. To score the productions of English children in Reception Year and Year 1 (5–6 years old), Caravolas, Hulme, and Snowling (2001) developed a scale on which each grapheme in a child’s spelling receives 0 to 4 points based on its proximity to the phoneme that it represents. Thus, ‹bak› for *back* gets the full number of points possible for this word because ‹k› may spell /k/ in English, even though a single vowel letter followed by ‹k› is very uncommon at the ends of English words. The spellings ‹bk› and ‹t› for *back* receive fewer points because one or more phonemes have been deleted, misidentified, or spelled with a letter that never represents the phoneme in English. We refer to such scoring methods as *phoneme-based nonbinary* measures. Other researchers, primarily those working with students in Grade 2 and above, have used what we call *letter-based nonbinary* measures. These measures consider the extent to which the child uses letters that are conventionally correct for the word in question, but the measures are not binary, as correctness is. For example, one letter-based system allots 1 point for using the correct first letter of a word, 1 point for the correct last letter, and 1 point for each correct two-letter sequence (Frisby, 2016). Still other researchers have used what Treiman, Kessler, Pollo, Byrne, and Olson (2016) called *mixed nonbinary* methods. These are scales that allot more points to correct spellings than to spellings in which all phonemes are spelled with letters that represent the phonemes in other words but that use options that are not correct for the word in question (e.g., ‹bak› for *back*).

Only a few studies have compared different scoring systems (Clemens, Oslund, Simmons, & Simmons, 2014; Frisby, 2016; Ritchey et al., 2010; Treiman et al., 2016), in part because using alternative scoring systems has often involved laborious scoring by hand. Treiman et al., in the only study to our knowledge to have compared different methods of scoring the spellings of children in early primary school as predictors of their later spelling performance, found some surprising results. In this study, the spellings of 374 US and Australian children at the end of kindergarten (mean age 6;2 [years;months]) were scored in eight different ways. The scoring methods correlated highly with one another, but letter-based methods were better predictors of performance on a standardized spelling test at the end of Grade 2 (mean age 8;1) than were phoneme-based or mixed methods. This result is surprising because the previously mentioned theories and findings (Ehri, 1986, 2005, 2014; Frith, 1985; Landerl & Wimmer, 2008) lead us to expect relatively poor performance for scoring systems that require use of orthographically correct letters such as ‹e› in *come* and that penalize errors such as ‹kap› for *camp* as much as errors such as ‹dap› for *camp*. Another surprising finding was that the mixed method that Treiman et al. examined was not the best predictor of later spelling performance. Mixed methods have been the most popular way of assessing young children's spelling in recent years, as Treiman et al. documented in a literature review, and one might therefore have expected that these methods would have risen to the top.

In an attempt to replicate and extend the surprising findings of Treiman et al. (2016), the present study examined children from a different country and different educational system. Whereas the US and Australian children in the earlier study had 8 or more months of kindergarten experience and an average age of 6;2 when their spelling was first tested, the English children studied here received their first spelling test when they had been in Reception Year for only 4 or 5 months and were on average 5;1. Children took the spelling test again near the end of Reception Year, at an average age of 5;7, and again around the middle of Year 1, at an average age of 6;1. The outcome measure, which was given around the middle of Year 2 (mean age 7;3), was the spelling subtest of the Wechsler Objective Reading Dimensions (WORD; Rust, Golombok, & Trickey, 1993). In this test, as in most standardized spelling tests, pupils spell a series of words that are presented in sentences and their responses are scored as correct or incorrect. We scored children's spellings at Times 1, 2, and 3 according to letter-based, phoneme-based, and mixed measures, asking whether some measures were better predictors of Time 4 performance than were others and whether the best predictors of Time 4 performance changed across the three earlier times. The spelling data that we analysed came from the previously mentioned study by Caravolas et al. (2001). We went beyond the analyses reported there by comparing a variety of spelling measures.

Our first scoring method was the traditional measure of correctness. A second method, which was also binary, distinguished between spellings in which all phonemes were represented in the correct order with a letter or letter group that may represent the phoneme in English (e.g., ‹bak› for *back*) and spellings in which this was not the case (e.g., ‹bt› for *back*). We refer to this method as measuring *phonological plausibility*, a term often used in past research (e.g., Hayes, Kessler, & Treiman, 2011; Jalil & Rickard Liow, 2008; Landerl & Wimmer, 2008). Phonologically plausible spellings reflect an ability to segment spoken words into phonemes and knowledge of correspondences between phonemes and letters, although not necessarily knowledge of which spellings of a phoneme are orthographically correct in specific cases. According to Landerl and Wimmer, phonological plausibility is an ideal scoring method to use with young children when predicting later spelling performance. The children in the study of Landerl and Wimmer had received 8 months of literacy instruction when their spelling was first tested, and the researchers argued that it would not be appropriate to expect orthographically correct spelling at this time. What is important, Landerl and Wimmer suggested, is the ability to translate the phonemes in a word into a phonologically plausible letter sequence, whether that sequence is orthographically correct or not. This ability can be measured by scoring spellings as phonologically plausible or implausible.

We also included three nonbinary measures that were based on the concept of string edit distance. This is a way of quantifying the distance between two strings of symbols by determining the minimum number of operations needed to transform one into the other (Levenshtein, 1965). Each operation—addition, deletion, or substitution of a unit—is assigned a penalty. The penalties are summed, so that poorer spellings receive higher error scores. The first method of this type, *letter distance*, was the distance between the child's spelling string and the conventional spelling. For example, ‹bak› for *back* differs from the correct spelling in that one letter was omitted. We set a penalty of 1 for this operation in our main analyses, so ‹bak› for *back* receives an error score of 1. Because ‹bck› and ‹ack› also omit one letter of the correct spelling, they receive the same penalty, even though ‹bck› and ‹ack› fail to represent one phoneme of the target word whereas ‹bak› represents all phonemes. A second measure based on edit distance, *phoneme distance*, was the distance between the child's spelling and the closest phonologically plausible spelling of the word. No changes are required to make ‹bak› phonologically plausible, so its error score is 0. Because ‹bakl› includes one extraneous unit, it receives an error score of 1 in our main analyses, 1 being the penalty set for inserting an extra element. The third measure based on edit distance, *AMPR* (automated measure of phoneme representation; Treiman & Kessler, 2004), was a more lenient phoneme-based system. The AMPR system takes account of such things as young children's tendency to spell /t/ before /r/ as ‹ch›, in line with its pronunciation in this context; ‹chre› for *tree* receives an error score of 0.

Our final two measures were the mixed nonbinary measures developed by Masterson and Apel (2010) and used in a number of recent studies (e.g., Bailey, Arciuli, & Stancliffe, 2017; Clemens et al., 2014; Kim, Puranik, & Al Otaiba, 2015; McNeill, Wolter, & Gillon, 2017; Werfel & Krimm, 2015). For *SSS-E* (Spelling Sensitivity Score–Elements), a child's spelling of a word is broken into elements, which for one-morpheme words like those of the present study are letters or groups of letters that spell a single phoneme. Each element is coded as correct (3 points), represented with a spelling that is legal, in that it spells that sound in similar contexts in other words, but that is not correct (2), represented with an illegal letter or letter group (1), or omitted (0), and scores are averaged across the elements of the word. The *SSS-W* (Spelling Sensitivity Score-Words) score is that of the lowest-scoring element in the word. A correct spelling thus receives 3 points, a spelling with one incorrect but legal element (e.g., ‹bak› for *back*) receives 2 points, a spelling with one or more illegal elements (e.g., ‹bam› for *back*) receives 1 point, and other spellings receive 0 points. One attractive feature of the SSS system is that its developers have made available a computer program to aid with scoring (Masterson & Apel, 2015).

Our primary research question was how the measures compared as predictors of later spelling performance. The research and theory we have discussed lead us to expect that phoneme-based measures would fare best. This should be especially true at Time 1, when the children were younger and had less school experience than the end-of-the-year US and Australian kindergartners in the study of Treiman et al. (2016). Measures that give full credit for phonologically plausible letters such as ‹k› in *camp* may best capture the ability to segment spoken words into phonemes and the knowledge of sound–letter correspondences that are important for spelling. We asked whether the predictive value of the different methods changed at Times 2 and 3, as children's spelling skills increased.

A secondary research question was about the influence of certain scoring decisions on the predictive value of the measures. Our main analyses penalized substitutions involving the mirror image letter forms ‹b›/‹d› and ‹p›/‹q› as much as other substitutions, but subsidiary analyses examined the effect of not penalizing these substitutions. Treiman et al. (2016) found that letter reversals at the end of kindergarten tended to be associated with poorer spelling in second grade. This result leads us to expect that, at the later time points of the present study, measures that did not penalize substitutions involving ‹b›/‹d› and ‹p›/‹q› would be poorer predictors of later spelling performance than measures that did penalize these substitutions. We also examined the effect on the letter and phoneme distance measures of using different penalty values for insertions, deletions, and substitutions.

Method

Participants

The spelling data were from the 136 children (64 female) in the study of Caravolas et al. (2001) who produced spellings at all the time points. These children were monolingual speakers of British English from York, England. The spelling test was given first in January and February of Reception Year when the children had a mean age of 5;1 (range 4;4–5;6). The test was administered again in June and July of Reception Year (Time 2, mean age 5;7) and in January and February of Year 1 (Time 3; mean age 6;1). Children took the standardized spelling test that served as the outcome measure in April of Year Two (Time 4, mean age 7;3).

Materials

The spelling test at Times 1, 2, and 3 contained 97 monosyllabic, monomorphemic words that represented common objects and actions. The words are listed in the appendix. For each word, a picture was prepared that depicted the concept that the word represented. To ensure that the words were appropriate for comparing letter-based and phoneme-based scoring, we estimated the probability that each word would be spelled correctly if a child segmented it into phonemes and spelled each phoneme using spellings that occur in other words without regard to within-word context or lexical correctness. This process has a high likelihood of yielding correct spellings of /b/ and /æ/ in *back*, for example, because ‹b› and ‹a› are very common spellings of these phonemes. The phoneme /k/ has a lower chance of being spelled correctly because ‹ck› is a less common spelling of /k/ than is either ‹c› or ‹k›. The spelling probability for *back*, which is derived by multiplying the probabilities of the individual correspondences, is thus fairly low. We obtained probabilities for correspondences by developing a set of phoneme–letter alignments for the words in the Reception Year list of the Children's Printed Word database (Masterson, Stuart, Dixon, & Lovejoy, 2010), treating all letters as if they spell some sound either by themselves or in combination with adjacent letters. For example, the correspondences for /m/ included ‹m› (as in *mud*), ‹mb› (*comb*), ‹me› (*come*), and ‹mm› (*hummed*). We calculated the probability of each correspondence, weighting the words in which they occurred by their frequencies of occurrence in the database. The average spelling probability for the words in our spelling test was .34 (*SD* = .27), suggesting that knowledge of which phoneme–letter correspondences were appropriate for particular contexts or particular words would be needed to spell these words correctly.

The spelling task at Time 4 was the spelling subtest of the WORD (Rust et al., 1993). It asks children to spell a series of words, graded in difficulty, that are presented in sentences. The test includes some challenging words, such as homonyms, contractions, and words with unusual sound-to-spelling correspondences. Children in Year 2 begin with the seventh word in the list and continue until they misspell six consecutive words. If the child misspells the first word that is presented, the examiner works backward until the child spells five consecutive words correctly or reaches the first item. Children receive one point for each correct spelling, and words before the first one spelled by a child are counted as correct. Responses that include reversals of mirror-image letters are scored as incorrect.

Procedure

For the spelling test at Times 1, 2, and 3, children were asked to name each picture and then write the word. Spellings were elicited from picture prompts in order to obtain spellings that were based on children's own phonological representations. If a child did not produce the intended label for a picture, the experimenter helped the child to do so. Because of the children's young age, they were given a strip of alphabet letters to which they were encouraged to refer while spelling. At Time 1, the spelling test was administered to pairs of children over the course of five sessions. At Times 2 and 3, children took the spelling test in groups of two to four over four sessions. The standardized spelling test at Time 4 was also administered to children in small groups. There was little scope for children influencing each other's spellings or pronunciations during the test sessions because children interacted primarily with the experimenter and worked at their own pace.

Scoring of Spellings at Times 1 to 3

Each child's productions were scored using the schemes described below. Table 1 shows how several spellings were scored using each scheme. Data were missing for 2.5% of trials (e.g., because a child did not attempt a word), and these trials were excluded from the analyses.

Correctness. We scored each spelling as conventionally correct or incorrect and calculated the proportion of each child's spellings that were correct.

Phonological plausibility. Each spelling was scored for whether it transcribed each phoneme in the proper sequence with a letter or group of letters that may be used to spell that phoneme, with no extra letters. We based this scoring on the list of correspondences that was described earlier. Spellings that fit pronunciation variants in Northern British dialect were accepted. We calculated the proportion of each child's spellings that were phonologically plausible.

Letter distance. We used the computer program Ponto (Kessler, 2009) to determine the number of deletions, additions, and substitutions needed to transform the child's spelling of each word into the conventional spelling. As in Treiman et al. (2016), a spelling was penalized by 1 point in our main analyses if it omitted a required unit, 1 point if it inserted an extraneous unit, and 1.4 points if one unit was substituted for another. (The choice of 1.4 penalty points for substitutions was motivated by the fact that this number approximates the Euclidean distance between omission of a plausible letter and the addition of a plausible one.) Letters were required to be in the correct sequence. We calculated the mean letter distance score across the words spelled by each child.

Phoneme distance. Using the same list of correspondences described earlier, we used the program Ponto (Kessler, 2009) to measure the distance of each child's spelling of each word from each of the phonologically plausible spellings. The penalties were the same as for the letter distance score and, as for the letter distance measure, we required letters to be in the correct sequence. For a child's spelling of each word, we used the best (smallest) distance score. We calculated the mean distance score across the words written by each child.

AMPR. Using the correspondences from Treiman and Kessler (2004), as modified for British English, we determined the number of additions needed to transform the child's spelling of each word into a phonologically plausible spelling. Because the system was designed for use with young children, letters were not required to be in the correct sequence and extraneous letters were not penalized. The scoring was done using Ponto (Kessler, 2009), and we calculated each child's average distance score.

SSS-W and SSS-E. We determined the correct and legal spellings of each element in each word of our list, using when possible the dictionary that is part of the computerized version of the Spelling Sensitivity System developed by Masterson and Apel (2015). We adjusted some of the dictionary codings to reflect our participants' dialect. All letters of a child's spelling production are associated with one or more elements of the standard spelling, in left to right order, except that final ‹e› can be associated with the preceding vowel. Each element is then assigned a score based on whether the letters associated with it are correct (3), legal but incorrect (2), illegal (1), or whether the element is not associated with a letter (0). The average across all elements is the word's SSS-E score. The word's SSS-W is the score of its lowest-scored element. Because there can be more than one way to associate a child's spelling with the word's elements, we decided that the most widely acceptable approach would be to give children the most credit possible by picking among multiple candidates an association that has the highest SSS-W. In case of ties, we selected from that set the one with the highest SSS-E. We wrote a computer program to carry out this policy, and the program and the full set of scores are available in the supplemental materials at http://spell.psychology.wustl.edu/MASSS/.[1]

Scoring at Time 4

We determined the standard score on the spelling subtest of the WORD (Rust et al., 1993), which is based on the child's age.

Results

Analyses Involving All Children

The leftmost columns of data in Table 2 show descriptive statistics for each spelling measure at Times 1 to 3. Performance improved significantly across Times 1 to 3, F(2,270) > 223, p < .001, η_p^2 > .62, for all measures according to one-way analyses of variance; recall that lower scores on letter distance, phoneme distance, and AMPR correspond to better performance.

Table 3 shows the Pearson correlation coefficients among the spelling measures at each time point. All of the correlations were statistically significant (p < .001 by one-tailed tests; we used one-tailed tests here because the direction of each correlation was hypothesized in advance). The mean standard score on the Time 4 spelling test was 102.20 (*SD* 16.80), and spelling at earlier times correlated significantly with Time 4 spelling regardless of which spelling measure was used (*p* < .001 by one-tailed tests).

For each time point, we carried out a series of tests to determine whether some of the earlier measures correlated significantly more highly with Time 4 spelling than others. We used the test of Steiger (1980), employing absolute values of correlation coefficients in this and other analyses that compared correlations. Because we compared 21 pairs of measures at each time point, we adjusted the *p* values using the Holm–Bonferroni method.

At Time 1, letter distance scoring appeared to have a small edge. None of the differences among the pairs of correlation coefficients was statistically significant using the Holm–Bonferroni procedure, however.

At Time 2, there was a tendency for correctness to be the best predictor of Time 4 performance. At this time point, the association between correctness and Time 4 spelling was significantly stronger than the association between the other binary measure, phonological plausibility, and Time 4 spelling (*p* < .0001). The other differences were not statistically significant after the Holm–Bonferroni correction.

At Time 3, correctness was significantly more highly associated with Time 4 spelling than were any of the other measures (*p* ≤ .001 for all comparisons). SSS-W appeared to be the next best predictor, performing significantly better than phonological plausibility, phoneme distance, AMPR, and SSS-E (*p* ≤ .001). The poorest predictor of Time 4 spelling was phoneme distance, which performed significantly less well than each of the other measures (*p* ≤ .001) except for AMPR, which also tended to perform poorly.

When we repeated the analyses without penalizing substitutions of the mirror-image forms ‹b›/‹d› and ‹p›/‹q›, the magnitude of the correlations between the spelling measures and the outcome measures did not change significantly at Time 1 or 2. At Time 3, the correlations between the spelling measures and the outcome measure were slightly (an average of .01 across the measures) weaker when mirror-image substitutions were not penalized than when they were penalized. The difference was statistically significant in the case of the SSS-E and SSSW-measures (*p* < .008 using the Holm–Bonferroni method to adjust for the fact that there were 7 tests at each time point).

Children Who Spelled No Words Correctly at Times 1 and 2

The results presented so far do not support the idea derived from previous theory and research that phoneme-based measures would predict later spelling better than would letter-based measures. However, it is possible that a superiority for phoneme-based measures would emerge among the poorest spellers in the study. To test this idea, we conducted separate analyses for those 86 children who spelled no words correctly at Time 1. These children had a mean age of 5;0 at Time 1 and a mean score of 97.93 on the Time 4 spelling test. Table 2 shows the descriptive statistics for the Time 1 spelling measures for this group, and Table 4 shows the correlations of the Time 1 measures (except for phonological plausibility, which was at floor) with Time 4 spelling. For this group, SSS-W did not correlate significantly with Time 4 spelling but the other measures did. Tables 2 and 4 also show the results for the 41 children who produced no correct spellings at Time 2 (mean age at Time 1 = 5;0; mean score on Time 4 spelling test = 90.41). For this group of children, SSS-E and SSS-W at Time 2 did not correlate significantly with Time 4 spelling but letter distance, phoneme distance, and AMPR did. For both groups, there was a tendency for phoneme distance to outperform the other measures as a predictor of later spelling, but the correlation of Time 4 spelling with phoneme distance did not significantly exceed the correlations of Time 4 spelling with letter distance or AMPR. We did not separately analyse the data of children who spelled no words correctly at Time 3 because there were only 12 such children. The results were very similar when we repeated the analyses without penalizing substitutions of mirror-image letters.

Letter Distance and Phoneme Distance Scoring Involving Different Penalties

In the analyses reported so far, we computed letter distance and phoneme distance using penalty values for insertions, deletions, and substitutions of 1, 1 and 1.4, respectively. Additional analyses were conducted to determine how use of different penalty values affected the correlations between letter distance and phoneme distance scoring at Time 1 and spelling performance at Time 4. We report these analyses for Time 1 only because it is at this time that these measures appeared to be most useful as predictors of Time 4 performance. Table 5 shows the correlations of Time 1 performance with Time 4 spelling using different penalty values, both for systems that penalized substitutions of mirror-image letters and those that did not. The correlations of Time 1 performance with Time 4 spelling were not significantly affected by whether the penalty value for substitutions was set at 1.4; 1, as suggested by the idea that a substitution is a single error; or 2, as suggested by the idea that a substitution is equivalent to an insertion plus a deletion. When we compared schemes that did not penalize deletions and schemes that did not penalize substitutions to schemes that allotted 1 penalty point for each type of transformation, the correlations with Time 4 performance declined significantly in magnitude ($p < .004$, two-tailed, for all comparisons). The small drops in the magnitudes of the correlations when insertions were not penalized were not statistically significant. The correlations were slightly larger in magnitude when substitutions of mirror-image letters were not penalized than when they were penalized, but the difference was statistically significant only for letter distance scoring that did not penalize deletions ($p = .008$, two-tailed).

Discussion

The goal in learning to spell, most researchers and educators would agree, is the ability to produce correct spellings of words with little mental effort. People who have achieved this goal can concentrate on higher levels of the writing process, and their readers will not be distracted by misspellings. Not surprisingly, then, standardized spelling tests typically measure spelling ability in terms of correctness. However, correctness may not be a good measure of the knowledge of beginning spellers. It penalizes phonologically plausible errors such as ‹mit› for *mitt* and ‹bangc› for *bank* and so may not provide a good indicator of the ability to analyze spoken words into phonemes and the knowledge of sound-to-spelling mappings that are thought to be critical for spelling success. Moreover, young children often perform at floor level when their spellings are scored for correctness.

We examined these issues by comparing the predictive power of conventional correctness and six alternative measures. One alternative is binary, like correctness, but it treats all phonologically plausible spellings alike whether or not they are orthographically correct. Three other alternatives, which are nonbinary, are based on the concept of edit distance. One of these, letter distance, assesses the distance between the letter string produced by a child and the correct spelling of the word. All omissions of a single letter are treated alike, for example, regardless of the phonological plausibility of the result. The other edit distance measures, phoneme distance and AMPR, assess failures to spell a phoneme, additions of letters or digraphs that do not represent a phoneme, and cases in which a child represents a phoneme with a letter or digraph that does not spell that phoneme in English. These phoneme-based measures do not penalize such things as use of ‹k› in the spelling of *camp*, as letter-based measures do. Our final two measures, SSS-E and SSS-W, are scales that give more credit to correct spellings than to incorrect but phonologically plausible alternatives. The English children in our study received the same spelling test around the middle of Reception Year, near the end of Reception Year, and around the middle of Year 1, and we scored their performance at each time point on each measure.

Whatever spelling measure we used, we found significant correlations between children's spelling performance at earlier times and performance on the outcome measure at Time 4 when we examined the results for the full group of children. These results suggest that there is some stability in the rate of spelling development in English children from at least the middle of Reception Year. Similarly, other studies have found stability from the second half of kindergarten to first or second grade in US and Australian children (McBride-Chang, 1998; Treiman et al., 2016).

As early as the end of Reception Year, Time 2, correctness was a significantly better predictor than was the phoneme-based binary measure, phonological plausibility—a measure that, according to Landerl and Wimmer (2008), should have performed quite well. Correctness was a significantly better predictor than each of the other measures at Time 3, around the middle of Year 1. The good performance of correctness as a predictor of later spelling performance is in some ways not surprising, because the outcome measure is also based on correctness. What is surprising is that correctness outperformed other measures even early in the development of spelling. It is often suggested that the main driver of spelling development until around 8 years of age is learning to analyze spoken words into phonemes and learning which letters may represent each phoneme (Ehri, 1986, 2005, 2014; Frith, 1985). It takes several years, according to these theories, for children to begin learning that some spellings of a phoneme are correct for some words but not others or for some contexts within words but not others.

The theories and research we have discussed (Ehri, 1986, 2005, 2014; Frith, 1985; Landerl & Wimmer, 2008) predict that phoneme-based scoring of young children's spelling should provide a good measure of the skills that are important during this period of development and should be a good predictor of later performance. The phoneme-based scoring methods that we used did correlate significantly with later performance. Importantly, however, there was no time point or no group of children for whom the phoneme-based methods emerged as significantly better predictors than the others. At Time 3, indeed, the phoneme distance measure seemed to be the poorest predictor. In the study of Treiman et al. (2016), too, the phoneme-based edit distance measure was a less good predictor of later spelling performance than was the letter-based edit distance measure. These findings fit with other evidence that even young children can attend to and remember visual orthographic features of words (Cassar & Treiman, 1997; Martinet, Valdois & Fayol, 2004; Wright & Ehri, 2007). It is possible that phoneme-based measures would be significantly better predictors of future performance than letter-based measures for children who are even less advanced than those studied here. The analyses that we conducted with the poorest spellers in the study provide a hint of such an effect, and further research is needed to examine this possibility.

Scales that allot more points for fully correct spellings than for phonologically plausible but legal errors and more points for spellings with more legal elements than for spellings with fewer legal elements have been widely used with young children. Surprisingly, given this popularity, the scoring system of this type that was included in the study of Treiman et al. (2016)—a scale that was adapted by Byrne and Fielding-Barnsley (1993) from the one introduced by Liberman, Rubin, Duquès, and Carlisle (1985)—did not rise to the top as a predictor of later spelling performance. The present study included two other measures in this category, SSS-E and SSS-W. The SSS-W measure was not a significant predictor of later spelling performance for children who spelled no words correctly at Time 1, and neither SSS-E nor SSS-W were significant predictors of later performance for children who spelled no words correctly at Time 2. Given that one motivation for the use of such scales has been to pick up meaningful differences among children who do not yet spell any words correctly, these findings are concerning. One contributor may be that the scores are rank values; it is far from certain that one correct letter (3 points in SSS) has three times the effect of an illegal letter (1 point) or one and a half times the effect of a legal letter (2).

The decision about whether to count reversals of the mirror-image forms ‹b›/‹d› and ‹p›/‹q› as correct did not have a large impact on the predictive value of the measures, although there were indications that reversals were a negative sign among older and more skilled spellers but not among younger and less skilled spellers. Not surprisingly, counting deletions and substitutions of required units as correct significantly lowered the predictive value of the letter distance and phoneme distance measures. The predictive value of these measures did not decline significantly, however, when insertions were scored as correct. Changes in how insertions were scored may have had a small impact in this study because our participants were less likely to insert letters that were not required than to make omissions or substitutions of required letters.

The goal of our study was to compare the ability of different spelling measures to predict later spelling performance. Finding a measure that predicts well is important for determining which children may go on to develop spelling difficulties and controlling for autoregressive effects in studies of the determinants of spelling performance (e.g., Caravolas et al., 2012), as well as for other purposes. Different choices of measures may be appropriate when the goal is to determine which linguistic and orthographic features to target in instruction for a particular child (Apel & Masterson, 2001; Ganske, 1999). For this purpose, it can be important to distinguish between spellings that omit phonemes (e.g., ‹bow› for *blow*) and those that do not (e.g., ‹blo› for *blow*), for example. It may also be helpful to use spelling tests that systematically sample words with different features, such as initial consonant clusters. A child who produces frequent spellings such as ‹bow› for *blow* may need instruction that targets the analysis of spoken clusters into smaller units.

The present study used Grade 2 performance on a standardized spelling test as the outcome measure. Such a test assesses the conventional spelling skills that are valued by society. Other ways of scoring performance on the Grade 2 test may have affected the predictive power of the earlier measures. We did not compare different methods of scoring the Grade 2 test here, and that could be done in future studies. It will also be important to examine languages other than English. French, for example, includes many words such as *fruit* ‘fruit’ that are spelled with a final letter that is not pronounced. Omission of the ‹t› of *fruit* is not penalized in phoneme-based scoring systems but is penalized in letter-based and mixed systems. Such omissions are common among young spellers of French (e.g., Sénéchal, Gingras, & L’Heureux, 2016), and future studies could compare different ways of scoring children’s early spelling of words like *fruit* as predictors of later spelling performance. In languages like Italian, which have simpler sound-to-spelling links, phoneme-based and letter-based scoring are more difficult to distinguish than they are in English or French. Even in these languages, however, there are some words that permit a distinction.

Overall, our results show that the scoring of children’s spellings for correctness allows for good prediction of their future spelling success and that this is true from a surprisingly early point in development. Before this point, nonbinary measures based on edit distance have promise. These measures can be computed using the freely available Web-based program used here (Kessler, 2009), and our results show that they predict later performance within a reasonable range of parameters. The letter distance measure is especially easy to use because decisions about the phonological forms of words and the plausible spellings of phonemes are not required. This is a reason to recommend it. Surprisingly, given their popularity in recent studies, the SSS-W and SSS-E measures did not fare very well as predictors of later spelling performance among children who did not yet spell words correctly.

References

Apel, K., & Masterson, J. J. (2001). Theory-guided spelling assessment and intervention. *Language Speech and Hearing Services in Schools*, *32*, 182–195. https://doi.org/10.1044/0161-1461(2001/017)

Apel, K., & Masterson, J. J. (2015). Comparing the spelling and reading abilities of students with cochlear implants and students with typical hearing. *Journal of Deaf Studies and Deaf Education*, *20*, 125–135. https://doi.org/10.1093/deafed/env002

Bailey, B., Arciuli, J., & Stancliffe, R. (2017). Effects of ABRACADABRA literacy instruction on children with autism spectrum disorder. *Scientific Studies of Reading*. https://doi.org/10.1037/edu0000138

Byrne, B., & Fielding-Barnsley, R. (1993). Evaluation of a program to teach phonemic awareness to young children: A 1-year follow-up. *Journal of Educational Psychology*, *85*, 104–111. https://doi.org/10.1037/0022-0663.85.1.104

Caravolas, M., Hulme, C., & Snowling, M. J. (2001). The foundations of spelling ability: Evidence from a 3-year longitudinal study. *Journal of Memory and Language*, *45*, 751–774. https://doi.org/10.1006/jmla.2000.2785

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

Cassar, M., & Treiman, R. (1997). The beginnings of orthographic knowledge: Children's knowledge of double letters in words. *Journal of Educational Psychology*, *89*, 631–644. https://doi.org/10.1037//0022-0663.89.4.631

Clemens, N. H., Oslund, E. L., Simmons, L. E., & Simmons, D. (2014). Assessing spelling in kindergarten: Further comparison of scoring metrics and their relation to reading skills. *Journal of School Psychology, 52*, 49–61. https://doi.org/10.1016/j.jsp.2013.12.005

Ehri, L. C. (1986). Sources of difficulty in learning to spell and read. In M. L. Wolraith & D. Routh (Eds.), *Advances in Developmental and Behavioral Pediatrics. Volume 7* (pp. 1211–195). Greenwich, CT: JAI Press.

Ehri, L. C. (2005). Development of sight word reading: Phases and findings. In M. J. Snowling & C. Hulme (Eds.), *Science of reading: A handbook* (pp. 135–154). Malden, MA: Blackwell.

Ehri, L. C. (2014). Orthographic mapping in the acquisition of sight word reading, spelling memory, and vocabulary learning. *Scientific Studies of Reading, 18,* 5–21. https://doi.org/10.1080/10888438.2013.819356

Frisby, C. (2016). An empirical comparison of the words spelled correctly and correct letter sequence scoring methods in third- and fourth-grade classrooms. *Journal of Applied School Psychology*, *32*, 101–121. https://doi.org/10.1080/15377903.2016.1151847

Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. Patterson, J. C. Marshall, & M. Coltheart (Eds.), *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading* (pp. 301–330). London, England: Lawrence Erlbaum Associates.

Ganske, K. (1999). The Developmental Spelling Analysis: A measure of orthographic knowledge. *Educational Assessment, 6,* 41–70. https://doi.org/10.1207/S15326977EA0601

Graham, S., Harris, K. R., & Chorzempa, B. F. (2002). Contribution of spelling instruction to the spelling, writing, and reading of poor spellers. *Journal of Educational Psychology, 94,* 669–686. https://doi.org/10.1037//0022-0663.94.4.669

Hayes, H., Kessler, B., & Treiman, R. (2011). Spelling of deaf children who use cochlear implants. *Scientific Studies of Reading*, *15*, 522–540. https://doi.org/10.1080/10888438.2010.528480

Jalil, S. B., & Rickard Liow, S. J. (2008). How does home language influence early spellings? Phonologically plausible errors of diglossic Malay children. *Applied Psycholinguistics*, *29*, 535–552. https://doi.org/10.1017/S0142716408080235

Kessler, B. (2009). Ponto [Computer software]. Available at http://spell.psychology.wustl.edu/ponto.

Kim, Y.-S., Puranik, C., & Al Otaiba, S. (2015). Developmental trajectories of writing skills in first grade. *Elementary School Journal*, *115*, 593–613. https://doi.org/10.1086/681971

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

Levenshtein, Vladimir I. (1965). Двоичные коды с исправлением выпадений, вставок и замещений символов [Binary codes for correcting deletions, insertions, and reversals of symbols]. *Doklady Akademii Nauk SSSR*, *163*, 845–848.

Liberman, I. Y., Rubin, H., Duquès, S., & Carlisle, J. (1985). Linguistic abilities and spelling proficiency in kindergarteners and adult poor spellers. In D. B. Gray & J. F. Kavanagh (Eds.), *Biobehavioral measures of dyslexia* (pp. 163–176). Parkton, MD: York Press.

Martinet, C., Valdois, S., & Fayol, M. (2004). Lexical orthographic knowledge develops from the beginning of literacy acquisition. *Cognition*, *91*, B11–B22. https://doi.org/10.1016/j.cognition.2003.09.002

Masterson, J. J., & Apel, K. (2010). The Spelling Sensitivity Score: Noting developmental changes in spelling knowledge. *Assessment for Effective Intervention*, *36*, 35–45. https://doi.org/10.1177/1534508410380039

Masterson, J., & Apel, K. (2015). Computerized Spelling Sensitivity System (CSSS) manual, vol. 1. https://www.missouristate.edu/assets/LLL/CSSS_Manual_2015.pdf

Masterson, J., Stuart, M., Dixon, M., & Lovejoy, S. (2010). Children's printed word database: Continuities and changes over time in children's early reading vocabulary. *British Journal of Psychology*, *101*, 221–242. https://doi.org/10.1348/000712608X371744

McBride-Chang, C. (1998). The development of invented spelling. *Early Education and Development, 9,* 147–160. https://doi.org/10.1207/s15566935eed0902_3

McNeill, B. C., Wolter, J., & Gillon, G. T. (2017). A comparison of the metalinguistic performance and spelling development of children with inconsistent speech sound disorder and their age-matched and reading-matched peers. *American Journal of Speech-Language Pathology, 26*, 456–468. https://doi.org/10.1002/em

Overby, M. S., Masterson, J. J., & Preston, J. L. (2015). Preliteracy speech sound production skill and linguistic characteristics of Grade 3 spellings: A study using the Templin Archive. *Journal of Speech, Language, and Hearing Research, 24*, 1–14. https://doi.org/10.1044/2015

Ritchey, K. D., Coker, D. L., & McCraw, S. B. (2010). A comparison of metrics for scoring beginning spelling. *Assessment for Effective Intervention, 35*, 78–88. https://doi.org/10.1177/1534508409336087

Rust, J., Golombok, S., & Trickey, G. (1993). *Wechsler Objective Reading Dimensions.* London, England: Psychological Corporation.

Sénéchal, M., Gingras, M., & L’Heureux, L. (2016). Modeling spelling acquisition: The effect of orthographic regularities on silent-letter representations. *Scientific Studies of Reading, 20,* 155–162. https://doi.org/10.1080/10888438.2015.1098650

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Treiman, R., & Kessler, B. (2004). The case of case: Children’s knowledge and use of upper- and lowercase letters. *Applied Psycholinguistics, 25,* 413–428. https://doi.org/10.1017/S0142716404001195

Treiman, R., Kessler, B., Pollo, T. C., Byrne, B., & Olson, R. K. (2016). Measures of kindergarten spelling and their relations to later spelling performance. *Scientific Studies of Reading*. https://doi.org/10.1080/10888438.2016.1186168

Werfel, K. L., & Krimm, H. (2015). Utility of the Spelling Sensitivity Score to analyze spellings of children with specific language impairment. *Australian Journal of Learning Difficulties, 20,* 39–53. https://doi.org/10.1080/19404158.2015.1047871

Wright, D.-M., & Ehri, L. C. (2007). Beginners remember orthography when they learn to read words: The case of doubled letters. *Applied Psycholinguistics*, *28*, 115–133. https://doi.org/10.1017/S0142716407070063

Footnote

[1] Because several recent studies have suggested that spellings that do not include at least two elements that are represented correctly or with a legal letter or digraph should be dropped from analyses using the SSS system (Apel & Masterson, 2015; Overby, Masterson, & Preston, 2015), we carried out additional analyses that did not include such spellings. The results, which are reported in the supplemental materials, show that eliminating low-quality spellings from SSS-E and SSS-W scoring did not usually yield higher correlations with Time 4 performance for all children as a group or children who produced no correct spellings at Time 1 or Time 2.

Table 1

Examples of Scoring Using Different Systems

Target word	Child's Spelling	Correctness (1=correct)	Phonological plausibility (1=plausible)	Letter distance	Phoneme distance	AMPR	SSS-E	SSS-W
‹back›	‹back›	1	1	0	0	0	3.00	3
‹back›	‹bac›	0	1	1 (1 deletion)	0	0	2.67	2
‹back›	‹b›	0	0	3 (3 deletions)	2 (2 deletions)	2	1.00	0
‹nut›	‹nrt›	0	0	1.4 (1 substitution)	1.4 (1 substitution)	1	2.33	1
‹tub›	‹tueb›	0	0	1 (1 insertion)	1 (1 insertion)	0	2.33	1
‹pen›	‹gyq›	0	0	4.2 (3 substitutions)	4.2 (3 substitutions)	3	1.00	1

Table 2

Means (and Standard Deviations) of Spelling Measures for Different Time Points and Groups of Children

Measure	Group and time point				
	Time 1, all children (*N* = 136)	Time 2, all children (*N* = 136)	Time 3, all children (*N* = 136)	Time 1, children with no correct spellings (*N* = 86)	Time 2, children with no correct spellings (*N* = 41)
Correctness	0.06 (0.13)	0.17 (0.21)	0.40 (0.24)	0.00 (0.00)	0.00 (0.00)
Phonological plausibility	0.11 (0.21)	0.29 (0.32)	0.60 (0.30)	0.00 (0.01)	0.00 (0.01)
Letter distance	3.17 (0.91)	2.50 (1.22)	1.53 (1.21)	3.66 (0.48)	3.82 (0.68)
Phoneme distance	2.45 (0.93)	1.75 (1.14)	0.85 (1.01)	2.97 (0.45)	3.03 (0.53)
AMPR	2.13 (0.89)	1.41 (0.99)	0.56 (0.67)	2.64 (0.39)	2.54 (0.39)
SSS-E	1.13 (0.69)	1.72 (0.77)	2.42 (0.49)	0.74 (0.32)	0.86 (0.34)
SSS-W	0.41 (0.61)	0.95 (0.82)	1.82 (0.64)	0.09 (0.18)	0.21 (0.30)

Table 3

Correlations of Measures at Each Time with One Another and Time 4 Spelling

		Phonological	Letter	Phoneme	AMPR	SSS-E	SSS-W	Time 4
Time	Measure	plausibility	distance	distance				spelling
1	Correctness	.98	-.84	-.85	-.85	.85	.94	.51
	Phonological plausibility		-.87	-.88	-.89	.88	.96	.48
	Letter distance			.99	.93	.-90	-.79	-.57
	Phoneme distance				.97	-.94	-.83	-.56
	AMPR					-.99	-.90	-.54
	SSS-E						.92	.53
	SSS-W							.46
2	Correctness	.97	-.87	-.86	-.87	.87	.94	.60
	Phonological plausibility		-.90	-.92	-.93	.92	.96	.52
	Letter distance			.99	.92	-.89	-.81	-.56
	Phoneme distance				.97	-.94	-.85	-.54
	AMPR					-.99	-.92	-.55
	SSS-E						.94	.54
	SSS-W							.54
3	Correctness	.93	-.86	-.80	-.82	.87*	.97	.75
	Phonological plausibility		-.93	-.92	-.93	.95	.96	.64
	Letter distance			.99	.95	-.95	-.85	-.61
	Phoneme distance				.97	-.96	-.82	-.55
	AMPR					-.99	-.87	-.58
	SSS-E						.92	.62
	SSS-W							.70

Note. *p* < .001, one-tailed, for all correlations.

Table 4

Correlations of Time 1 Measures with One Another and Time 4 Spelling for Children with No Correct Spellings at Time 1 (N = 86, Above Diagonal) and Correlations of Time 2 Measures with One Another and Time 4 Spelling for Children with No Correct Spellings at Time 2 (N = 41, Below Diagonal)

	Letter distance	Phoneme distance	AMPR	SSS-E	SSS-W	Time 4 spelling
Letter distance		.97***	.73***	-.55***	.38	-.46***
Phoneme distance	.92***		.84***	-.69***	.21	-.49***
AMPR	.28	.57***		-.96***	-.33***	-.48***
SSS-E	.01	-.32*	-.94***		.54***	.43***
SSS-W	.72	.44	-.43**	.69***		.01
Time 4 spelling	-.42**	-.48***	-.34*	.22	-.12	

$^{*}p < .05$, one-tailed. $^{**}p < .01$, one-tailed. $^{***}p < .001$, one-tailed

Table 5

Correlation of Letter Distance and Phoneme Distance Scoring of Time 1 Spelling with Time 4 Spelling Using Different Penalty Points for Insertions, Deletions, and Substitutions and Penalizing or Not Penalizing Substitutions of Mirror-image Letter Forms

Insertion penalty	Deletion penalty	Substitution penalty	Mirror-image substitutions penalized	Letter distance	Phoneme distance
1	1	1.4	yes	-.57***	-.56***
1	1	1	yes	-.57***	-.55***
1	1	2	yes	-.55***	-.55***
0	1	1	yes	-.56***	-.54***
1	0	1	yes	-.22*	-.31***
1	1	0	yes	-.44***	-.43***
1	1	1.4	no	-.58***	-.56***
1	1	1	no	-.58***	-.55***
1	1	2	no	-.56***	-.56***
0	1	1	no	-.56***	-.54***
1	0	1	no	-.24**	-.31***
1	1	0	no	-.44***	-.43***

$^{*}p < .05$, one-tailed. $^{**}p < .01$, one-tailed. $^{***}p < .001$, one-tailed

Appendix: Words in spelling task at Times 1, 2 and 3

back, bag, ball, bank, bath, bed, bell, belt, bone, book, boot, bowl, brick, broom, bump, bun, cake, camp, cap, car, card, cave, chick, coat, cod, cold, crib, crown, cup, dog, door, dot, dove, dream, dress, drink, dwarf, fan, fast, fist, foot, fork, gift, glass, gold, gull, gum, ham, heart, hen, hunt, hut, jam, jar, jet, lamb, laugh, leap, leg, lip, man, mask, mat, milk, mill, mitt, mud, nest, net, nut, pan, peach, pear, pen, pig, pin, pink, pond, rain, root, rug, sad, sand, soup, sun, tart, ten, tent, train, tree, trunk, tub, vase, veil, vest, wood, yawn