Evaluating the Spelling and Handwriting Legibility Test (SaHLT):

Downing, Cameron; Caravolas, Marketa

Reading and Writing

DOI:
10.1007/s11145-022-10402-2

Published: 01/01/2024

Peer reviewed version

Evaluating the Spelling and Handwriting Legibility Test (SaHLT): A tool for the Concurrent Assessment of Spelling and Handwriting

Cameron Downing[1,2] (Orcid ID: 0000-0001-7350-4048) and Markéta Caravolas[2,3] (Orcid ID: 0000-0003-4171-482X)

[1]School of Psychology and Therapeutic Studies, Leeds Trinity University, Leeds, United Kingdom
[2]Miles Dyslexia Centre, Bangor University, Bangor, United Kingdom
[3]School of Human Behavioural Sciences, Bangor University, Bangor, United Kingdom

Corresponding author:

Dr Cameron Downing (c.downing@leedstrinity.ac.uk)
School of Psychology and Therapeutic Studies, Faculty of Social and Health Sciences
Leeds Trinity University
Brownberrie Lane
Horsforth
Leeds
LS18 5HD

Author contributions:

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by C. Downing and M. Caravolas. The first draft of the manuscript was written by C. Downing and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We wish to thank the children, parents, and schools who took part in the projects which led to this publication. We also wish to thank Caspar Wynne, for his assistance with developing and scoring using this measure.

Funding Source

Much of this work was funded by the Waterloo Foundation (grant number: 1939-3205) and we thank them for their support of this work.

Abstract

Spelling and handwriting are related skills which are critical for writing but are typically assessed separately. Doing so makes it more difficult to understand their respective development. We describe the creation and evaluation of a tool for their concurrent assessment: the Spelling and Handwriting Legibility Test (SaHLT). We examined whether (a) sentence spelling and handwriting legibility could be reliable and valid, independent measures of English spelling and handwriting legibility and (b) whether spelling and handwriting legibility can be measured concurrently. A total of 1,461 primary-aged children (M_{age} = 9.14 years-old, *SD* = 12.80) completed the SaHLT and background tests. The SaHLT was a sensitive, reliable, and valid measure of spelling and handwriting. Multi-group factor analyses revealed the test to be a robust concurrent measure. The SaHLT offers a cost and time saving method of measuring two key skills of writing. This is important for assessments in practice and for furthering our understanding of the relationship between spelling and handwriting.

Key words: Spelling, handwriting legibility, literacy, development

Statements and Declarations

The authors declare no competing interests. This work was funded by the Waterloo Foundation. This study was conducted in accordance with the British Psychological Society Code of Ethics and Conduct and received approval from [withheld for review].

The abilities to spell and to write legibly are foundational literacy skills. They set the stage for the development of the complex, higher-level skills of text writing, so critical for educational success (Berninger & Winn, 2006; Olive, 2014). Impairments in one or both skills are related to broader writing difficulties in children with developmental disorders such as dyslexia (e.g., Sumner et al., 2014), developmental coordination disorder (e.g., Prunty & Barnett, 2017), and attention deficit hyperactivity disorder (e.g., Adi-Japha et al., 2007). While spelling and handwriting are under the control of different domains (i.e., spelling - primarily cognitive; handwriting – primarily perceptuomotor), emerging research evidence suggests that they are inter-related in typically developing children (e.g., Afonso et al., 2018; Caravolas et al., 2020; Kandel et al., 2017) and in children with developmental disorders such as dyslexia (e.g., Kandel et al., 2017; Suárez-Coalla et al., 2020; Sumner et al., 2014). Yet, the two skills are typically assessed by separate measures that are designed to focus only on each skill in isolation. This practice makes it more difficult, time consuming, and costly to gauge the spelling and handwriting profiles within individuals and to understand the nature of their potential difficulties across the two domains. In the present paper, we report a study that concurrently investigated the development of spelling ability and handwriting legibility, and the nature of the relationship between them, in primary-school-aged children using a single measurement tool, the Spelling and Handwriting Legibility Test (SaHLT). In what follows, we summarise what is known about the development of spelling and handwriting ability, and then briefly review existing approaches to assessing spelling accuracy and handwriting legibility before turning to the development of the SaHLT.

Spelling Development and its Assessment

Learning to spell in alphabetic orthographies is a complex, long-term process that is founded on three early-acquired cognitive skills, namely the ability to consciously analyse and manipulate speech at the level of phonemes (phoneme awareness), knowledge of the letters of the alphabet, and the ability to learn the correspondences between phonemes and graphemes (i.e. the minimal letter strings used in correspondence with phonemes; Caravolas et al., 2012; Kessler & Treiman, 2003). With these basic skills, children can begin to learn to represent in graphic form the phonological (speech sound) structure, and eventually, with the benefits of growing reading skills and psycholinguistic insights, to produce correct spellings of the words they attempt to write (Caravolas et al., 2001; Nunes & Bryant, 2009; Treiman & Kessler, 2014).

The English orthography, which is the focus of the present study, is phonographemically less consistent than most other alphabetic orthographies (Caravolas, 2022; Caravolas & Kessler, 2016). Many children master this skill of *phonological recoding* by the second year of schooling, but measurable individual differences in phoneme-grapheme encoding skills persist among English-speaking children beyond this point (Caravolas et al., 2001; Caravolas & Kessler, 2016). Children with spelling difficulties may continue to struggle to represent attributes of written words that are more complex, such as consonant clusters, complex vowels (diphthongs), unstressed syllables, and polysyllabic words (e.g., Treiman, 2017). For this reason, tests assessing children's skills often contain words varying in terms of such attributes (Caravolas, 2004; Treiman, 1993).

From a psycholinguistic perspective, the inconsistent and complex nature of the English orthography requires that learners have not only good phonological skills, but also some knowledge about the morpho-phonological basis of the system (e.g., Nunes & Bryant, 2009). Moreover, learners need to be sensitive to the distributional properties of the letters and letter patterns occurring in printed words, many of which have no clear connection to phonology or morphology, but instead reflect archaic or foreign pronunciations (Treiman & Kessler, 2014). Therefore, conventional spelling competence requires the integration of higher-order knowledge about the morphological (Nunes et al., 1997; Treiman et al., 1994) and orthographic (Cassar & Treiman, 1997; Calhoon & Masterson, 2011) structure of words as well. Over the primary years, English children learn specific contextualized spelling patterns, such as contexts for short and long vowel spellings, consonant doubling, and letter sequencing regularities that reflect the morpho-phonemic as well as the historical nature of the spelling system, both through explicit spelling instruction, and to some extent, through incidental exposure to print during reading practice.

In the English tradition, spelling ability is usually measured in one of three test administration methods: by producing spellings of single words to dictation, by spellings obtained from samples of children's free writing, and by the recognition of correctly spelled words from incorrectly spelled alternatives. The first of these methods is arguably the most frequently used, with numerous reliable, standardised, single-word spelling tests being available for clinical and educational use (see Calhoon & Masterson, 2011). Typically, the tests are graded in difficulty spanning a broad age range, and item selection is based on spelling patterns that present a learning challenge because they do not adhere to basic (canonical) sound-letter mappings that can be derived by "sounding out" the words' phonemes. Measures are usually scored for overall accuracy, and as such they tend to be highly reliable and sensitive to age/schooling levels but provide limited insights into the causes of children's misspellings. More recent approaches (e.g., Caravolas et al., 2018; Daffern & Ramful, 2020; Masterson & Apel, 2010; Treiman & Bourassa, 2000) select test items that reflect the variety of linguistic and orthographic sources of sound-letter inconsistency (e.g., graphotactic, morphophonological, etymological). This approach is traditional in many languages with more consistent orthographies (e.g., Czech: Caravolas & Volín, 2001, 2005; Slovak: Caravolas et al., 2008; Greek: Sideridis et al., 2008) and it provides richer information about children's command of the psycholinguistic and orthographic constraints on the written language they are learning. A variant of this approach, where words graded in difficulty were embedded in dictated sentences, was also adopted in the SaHLT tool we describe here.

While the optimal method of spelling assessment is likely to depend on the aims of the test and the context of the learner, the single-word format is adequate for estimating general spelling skills in English, assuming well-selected word lists, because the language has relatively low morphological (especially inflectional) complexity, as compared to many languages. Consequently, individual word spellings are not so affected by grammatical agreement among neighbouring words in phrases and sentences. In contrast, in many languages with alphabetic orthographies *and* with rich and more complex morphology and grammar, word spellings may alter depending on the syntactic/grammatical context of neighbouring words, and accordingly, spelling tests often comprise short written texts, or phrases containing the target words and their disambiguating context – which itself may be phonographemically inconsistent (e.g., Bosse et al., 2021; Caravolas et al., 2008). Nevertheless, the encoding of morphological information in word spellings and the existence of heterographic homophones in the English orthography typically require aurally presented carrier sentences to provide context in single word spelling tests. Arguably, tests requiring the writing of full sentences better approximate a typical writing situation and, as such, provide an ecologically valid assessment of spelling (Department for Education, 2014). However, sentence writing methods for assessing English spelling are seldom used in standardised tests while single word dictation is the primary assessment method (Treiman et al., 2019).

The spelling measure discussed in this study – named the Spelling and Handwriting Legibility Test (SaHLT) – was adapted from a test originally created for cross-linguistic comparison of English and the highly inflected Czech language (Caravolas & Volín, 2001; Caravolas et al., 2005). It adopts the idea of graded word lists of the English tradition and embeds the words within disambiguating linguistic (syntactic) contexts, and requires the whole sentences to be written, a format typical of the Czech tradition. The test consists of ten sentences, containing 62 words in total, which accrue in spelling difficulty both in phonological recoding demands (word length, syllable structure – consonant clusters, complex vowels) and with letter-sound inconsistencies due to graphotactic constraints on letter sequences, morphological patterns, and etymological exceptions.

Handwriting Legibility Development and its Assessment

Handwriting ability is typically considered in terms of fluency and legibility. Handwriting fluency is the ability to produce written script at speed without undue hesitation or effort. Legibility pertains to the ability to produce written scripts such that they are easily decipherable to the reader; albeit norms for well-formed, legible scripts are to some extent culturally determined (Kapoor & Saini, 2017). Whereas handwriting legibility is difficult to assess objectively, handwriting fluency is relatively easy to measure by recording the number of units (e.g., letters or words) written within a specified time frame (e.g., Alamargot & Morin, 2015; Pontart et al., 2013). The relative ease in measuring handwriting fluency using such tools as the Detailed Assessment of Speed of Handwriting (Barnett et al., 2007) has led to it being more extensively studied and tested in the classroom and clinic (Lambert & Quémart, 2019). However, fluent handwriting requires sufficiently consolidated representations of letters and words (Thibon et al., 2018) and so fluency measures may not be reliable for less able (younger and impaired) learners who have yet to consolidate representations. In contrast, handwriting legibility reflects the quality of these representations, at least in part (e.g., Pritchard et al., 2021), and so may be more reliable and appropriate for less skilled writers. Handwriting fluency and legibility are separable skills with different developmental trajectories (Authors, submitted, Gosse et al., 2021; Graham et al., 1998). The ability to write legibly is critical for accessing education and remains an important component of curricula (Cornhill & Case-Smith, 1996; Department for Education, 2014). It is therefore important to measure handwriting legibility in addition to handwriting fluency.

Relatively fewer investigations have focused on handwriting legibility than on handwriting fluency. Legibility measures can be categorized as *global* or *analytic*. Global measures are those where criteria are applied to the written product (e.g., a multi-sentence text) as a whole, whereas analytic measures typically apply criteria to specific attributes of legibility, such as formation, size, slanting, spacing, and straightness of letters, but different scales operationalise these attributes differently. In their review, Rosenblum et al. (2003) have argued that analytic measures have become more popular than global measures (although, see Barnett et al., 2018 for a recent example of a global handwriting measure). Despite many measures broadly agreeing on the attributes of legibility, the theoretical rationale for selecting these attributes remains unclear. Furthermore, Rosenblum et al. (2003) identified several issues with current analytic measures of handwriting legibility. These included (a) a lack of clarity in the scope of measures (i.e., for what purpose the measure could be used, be it screening and assessment of handwriting difficulties and/or tracking typical development), (b) practical administration of the measure (i.e., the ease of use in classroom, clinic, and research settings), and (c) how well the measure has been evaluated in terms of its psychometric properties. Indeed, of the 13 analytic measures reviewed, only 2 scales had undergone a comprehensive evaluation of their reliability, validity, and developmental sensitivity.

The handwriting scale developed by Ziviani and Elkins (1984) measured legibility by asking children aged 8 to 12 years to *copy* letter-like symbols, alphabet letters, and words. Subsequently, raters scored the productions on criteria including formation, spacing, alignment, and size. The four criteria were measured using an overlay as a scoring aid, although the operationalization of the scale was not clear. Nevertheless, the authors showed that their scale was sensitive to age-related changes in legibility and undertook efforts to evaluate the reliability and validity of the measure. Whilst inter-rater reliability of non-linguistic (letter-like) symbols (r = .88 - .98) fell above the authors' own desirable cut-off (r = .8), the inter-rater reliability of linguistic items (letters and words) – and arguably most ecologically valid measures of legibility – fell below the threshold (.69–.76). Similarly, test-retest reliability (measured one week apart) overall was below the authors' threshold (.62–.84). Ziviani and Elkins (1984) attempted to demonstrate criterion validity by correlating scores of the criteria on different items. These correlations were moderate to large (r = .46–.76), but this method does not demonstrate criterion validity, rather, it further demonstrates the reliability of the measure. Finally, principal component analysis did verify that the criteria measured different components of handwriting legibility and the scale was sensitive to developmental changes. To our knowledge, there has been no further work to improve the scale's psychometric properties.

A second well-evaluated scale is a Dutch language assessment of handwriting: Beknopte Beoordelingsmethode voor Kinder Handschriften (BHK; Concise Assessment Methods of Children's Handwriting; Hamstra-Bletz et al, 1987). The scale was designed to screen for handwriting difficulties (dysgraphia) amongst children aged 5 to 9 years. Children copy a passage for 5 minutes, which is then scored using 13 criteria which are operationalised through a mixture of Likert type and yes/no responses. This measure has been well evaluated with the authors and others demonstrating its developmental sensitivity (Hamstra-Bletz & Blöte, 1990). Evaluation of the scale's reliability revealed interrater reliability (r = .76–.89) and internal consistency, although relatively low (α = .52), to be acceptable (Hamstra-Bletz et al., 1987). Similarly, the BHK was found to correlate well with teacher evaluations of children's handwriting quality (r = .78) suggesting convergent validity and discriminant validity in differentiating between children with and without dysgraphia (Smits-Englesman et al., 2001). However, a major, practical drawback of the BHK is the extensive time required for training in test administration and scoring, as pointed out by the authors.

To address the practical issues of the BHK, Smits-Engelsman et al. (2005) developed the Systematische Opsporing van Schrijfmotorische problemen (SOS) by identifying and reformulating the six most discriminating items from the BHK. The SOS has been found to be developmentally sensitive for children aged 7-to-11 years, with excellent interrater reliability (ICC = .77) and moderate test-retest reliability (ICC = .69; Van Waelvelde et al., 2012). SOS scores correlate highly with BHK scores (*r* = .70) suggesting good convergent validity and the measure also discriminates between children with and without special education needs (SEN), suggesting good discriminant validity (Van Waelvelde et al., 2012). Whilst both the BHK and SOS seem to be well evaluated for use in Dutch, it is not clear how applicable the measures are to English, which is linguistically and educationally different.

There is considerable variability in handwriting norms across languages and cultures (Kapoor & Saini, 2017). The root of these differences reflects factors such as the perceived educational value and techniques employed in handwriting instruction in different cultures. In the U.K. and U.S., there are considerable variations (within country) of the frequency, duration, and methods of instruction of handwriting (Graham et al., 2008; Stainthorp et al., 2006). However, in other countries, such as the Czech Republic, a highly consistent approach is taken, whereby all children are taught the same cursive script (Bartošová et al., 2012; Kučerová & Kucharská, 2018). Such cross-cultural differences in approaches to instruction align withlarge handwriting style and legibility variability in (e.g., cursivity and specific slant of the script) in countries where inconsistent approaches are taken (e.g., the U.K. and US; Asher, 2006), and greater uniformity in others that adhere to a consistent approach (e.g., the Czech Republic; Bartošová et al., 2012). Consequently, the cross-cultural generality of handwriting tests be limited, be they designed for a particular style (e.g., Ziviani & Elkins,

1984) or test type. For example, measuring the legibility of cursive writing is likely to be less valued in countries allowing flexible (or variable) approaches to handwriting instruction, and where cursive handwriting may not be taught or used by at all, than in countries where this is the norm (Kučerová & Kucharská, 2018). It is important that handwriting legibility assessments have ecological validity for the populations for which they are intended; at the same time, however, an interesting goal should be to develop scales evaluating components of handwriting which generalize for direct comparisons across script types, or at least within script families, as in alphabetic orthographies using Latin letters, for example. However, to date, no such attempt has been made.

Unfortunately, since Rosenblum et al.'s (2003) review, – to our knowledge – there has been little development or psychometric evaluation of analytic measures capable of measuring handwriting legibility over a range of ages amongst primary school children. The ten-sentence format of the SaHLT lends itself well to the assessment of handwriting legibility because it enables insights into children's handwriting skills in a context that is more akin to typical writing activities than do single word tests. Moreover, it allows assessment of handwriting at the inter- as well as the intra-word levels.

The SaHLT also offers a tool enabling the further understanding of the concurrent relationship between spelling and handwriting. Theoretically, van Galen's (1991) psychomotor model makes some predictions about the relationship between spelling and handwriting production with spelling processes being assumed to cascade down, exerting influence on the motor processes associated with handwriting. Experimental work examining the effects of spelling complexity and spelling ability on the dynamics or fluency of handwriting production (Kandel & Perret, 2015; Kandel et al., 2017; Suárez-Coella et al., 2020) lends support to the view that spelling processes exert some control over handwriting *fluency* processes. Only a few studies have examined the nature of the relationship between spelling and handwriting *legibility*. Caravolas et al. (2020) found spelling ability to be the strongest predictor of handwriting legibility over and above that of schooling experience and reading ability. Pritchard et al. (2021) examined this relationship in younger children and took the view that the quality of motoric representations realised in handwriting legibility holds an important role in later spelling development. This prediction is in the opposite direction of the prediction from van Galen's (1991) model. The authors found a small but significant relationship between handwriting legibility at 6 years and word spelling 6 months later. Gosse et al. (2021) examined the longitudinal relations between spelling, handwriting fluency, and legibility. Using an experimental task in which spelling accuracy, handwriting fluency, and handwriting legibility were derived from single words, the authors found that spelling accuracy was related to the development of fluency, yet no association between spelling accuracy and legibility was found. To date then, the little work on the relationship between spelling and handwriting reflects a lack of consensus on the relationship between the constructs. We argue that this, in part, stems from methodological differences in the measurement of spelling and handwriting. As such, given the importance of both skills in typical and atypical writing development, with the Spelling and Handwriting Legibility Test (SaHLT), we sought to develop and evaluate a concurrent measure of spelling and handwriting which can be used in the classroom, clinic, and in the lab.

Rationale and Description of the SaHLT

In the development of the SaHLT, we were motivated to further the theoretical and empirical understanding of the cognitive foundations of spelling knowledge and of motor processes in primary-school-aged children. Here, we provide a rationale and description of how the SaHLT was developed. We also provide a more detailed rationale for the handwriting subtest of the SaHLT in Online Resource 1.

In line with current understanding of spelling development, we selected words graded in difficulty, reflecting the range of spelling patterns that are taught in the British English curriculum over the primary years. The resulting measure of ten sentences (62 words) is demonstrably sensitive for use with children aged 7 to 12 years of age (Class Years 2 to 7) (Caravolas et al., 2005). The sentences range in length from 4 to 8 words, and word length varies from 1 to 9 graphemes (i.e., letters or letter groups corresponding to a single sound or phoneme). As noted earlier, the words and word structures were selected in a graded manner to include units of varying complexity in terms of phonological, morphological, lexical, and orthographic structure. The choice of targeted spelling units making up the word and sentence structures was informed by level-appropriate national curriculum guidelines for Key Stages 1 and 2 (Department for Education, 2014). Thus, the earlier items are shorter and easier from a spelling and grammatical point of view, while the later items include more difficult and longer words. Punctuation is also assessed such that sentence-initial words must be capitalised, and the apostrophe is required to mark contraction (<won't>) and possession (<musician's>). In addition, phonological complexity of the words was manipulated in terms of number of phonemes, number of syllables, and consonant clusters throughout the test. Phonologically, the words ranged from the single phoneme/single syllable word, through relatively simple one- or two-syllable words containing only two to four phonemes, to rather complex, multisyllabic words like <performance> - /pəfoːmənʦ/.

The handwriting subtest of the SaHLT was also developed on theoretical and empirical understanding of handwriting processes in children and adults (see Online Resource 1). By measuring characteristics of handwriting which reflect its underlying processes it is possible to identify individual differences in handwriting legibility (Graham et al., 2006), the latter possibly having literacy and/or motor difficulties (Downing, 2018). With this in mind, we selected four legibility dimensions – letter formation, letter spacing, word spacing, and line alignment – which tap the perceptual-motor processes used in children's handwriting production. Letter formation measures presumably rely heavily on letter representations and their associated motor programmes, and they assess the accuracy of each letter's motor programme. Specifically, knowledge of the letter is measured by the letter's form, orientation, and consistency of its angle and size. letter spacing and word spacing dimensions reflect the visuospatial attributes at the inter-letter and inter-word levels by measuring the degree and consistency of the spacing. Finally, line alignment taps both visuospatial and motor control processes. It follows that visuospatial planning is required for the child to position writing on the line, whilst motor control processes contribute to planning, initiating, and maintaining the letters on the line. Therefore, letters that float above/cross below the line may indicate difficulties in visuospatial planning and/or motor control.

The handwriting measure is a rating assessment of legibility applied to the child's written responses on the spelling test. The ratings are made at the sentence level on four dimensions of legibility using a Likert-scale from one (very illegible) to five (highly legible). Thus, ten ratings are generated for each dimension (one set per sentence). In line with some other analytic legibility tests (e.g., Ziviani & Elkins, 1984), we chose to develop an analytic scale with theoretically driven dimension choices to capture specific processes used in handwriting production at the letter and word level. The characteristics of our dimensions were developed by first deconstructing the legibility characteristics of the underlying handwriting processes (e.g., Graham et al., 2006; van Galen, 1991). For example, a letter that is well formed and recognisable suggests a good representation of its specific allographic form (e.g., lower case, cursive), as well as the ability to correctly execute its motor programme. Taking this task analysis approach, we iteratively refined the characteristics and component features of each of the four dimensions. The first and second author then independently scored several responses using the defined characteristics before revising the definitions over several sessions to improve clarity. This led to the creation of four dimensions: letter formation which measures the how well-formed letters within words are, letter spacing which measures the appropriateness of spacing between letters within words, word spacing which measures the appropriateness of the spacing between words within sentences, and line alignment which measures the degree to which the written output sits on the line.

In some cases, we combined characteristics of underlying processes – which are assumed to be different – into one dimension. Specifically, in adults, sizing is assumed to be an independent process to allograph selection (van Galen, 1991). However, in children, producing letters that are proportionately larger (<red>) or smaller (<red>) than other letters within a word (sizing errors) could reflect either poor letter formation knowledge, unstable/unconsolidated representations of specific allographs, problems at the motor programming level, or poor motor control. During the initial scoring, we found that children who made sizing errors often did so in a consistent manner for specific letters across words and sentences within their response. Therefore, these errors were more likely the result of poor letter formation knowledge and so letter sizing was considered a feature of the Letter Form dimension.

The Current Study

In this study we aim to investigate the concurrent relationship between spelling and handwriting legibility over four class year groups of mid-primary-school children. Specifically, we sought to first examine whether a sentence spelling task would be a sensitive, reliable, and valid measure of spelling ability and would correlate well with a widely used single word measure of spelling. Second, in view of the scant evidence of robust, concurrent measures of spelling and handwriting, we sought to evaluate whether it was possible to measure spelling *and* handwriting legibility concurrently in a reliable and valid way. Finally, we sought to investigate whether the concurrent relationship between spelling and handwriting legibility altered over a period of rapid development of both spelling and handwriting legibility (Gosse et al., 2021; McCutcheon, 1996). We hypothesize that they are separate, but related constructs. Furthermore, we predict that the relationship between spelling and handwriting legibility will become weaker in older children as spelling and handwriting skills become automatised.

Method

Participants

A total of 1,461 children completed the SaHLT as part of three projects between 2015 and 2020 (please note that some of the monolingual participants in Caravolas et al., 2020) are also included in this sample). This sample size was determined to be sufficient for the analysis techniques proposed below. The children attended school Class Years 3–6 and came from schools located in North Wales (see Table 1). Relatively fewer children in Class Year 6 took part as they completed the measures used in this sample as part of a smaller project. All schools taught predominantly through the medium of English and used English for day-to-day communication; Welsh was formally taught as a second language. Of this sample, 94% spoke English as their home language, 0.3% spoke Welsh, and the remaining 5.7% spoke a variety of other languages. A brief survey revealed that the group was predominantly right-handed (88%), with 11.8% being left-handed and 0.2% being ambidextrous. The proportion of children on free school meals – a proxy for SES – in the participating schools ranged from 2.3% to 35.3%. The average number of children on school meals in our sample (18.9%) was very similar to the average of the local area (18.8%) and of Wales (20.1%; Welsh Government, 2018).

Participants completed the SaHLT along with measures of spelling, and teachers completed the handwriting proficiency screening questionnaire (HPSQ; see next section). For all participants, head teachers provided full written informed consent, while parents provided opt-out informed consent. This consenting procedure was used as testing was carried out in the classroom and constituted typical classroom activity. This study was conducted in accordance with the British Psychological Society Code of Ethics and Conduct.

[Table 1 here]

Measures

Spelling and Handwriting Legibility Test (SaHLT)

We offer a comprehensive description of this measure in the introduction and Online Resource 1 and so we only briefly describe it here. The SaHLT comprises spelling and handwriting legibility which can be administered to children in groups or individually. The spelling subtest includes ten sentences dictated aloud, and children are asked to write the sentence exactly as they hear them on lined paper in their normal handwriting. Each sentence was repeated at regular intervals until the last child had finished writing. Accuracy was measured by awarding one point for every correctly spelled word. The maximum score was 62.

Handwriting legibility was scored separately, and without regard for spelling accuracy, by a trained research assistant or teacher. The handwriting legibility scoring criteria were applied at the sentence level to the written productions. Each of the ten sentences was scored according to the criteria of each dimension (letter formation, letter spacing, word spacing, and line alignment) on a five-point Likert scale ranging from 1 to 5 (very illegible to highly legible). The scale was applied aggregating over the relevant units (letters, words, respectively) for each of the dimensions in each sentence. An overall legibility score for each sentence was obtained by summing across the four-dimension scores (possible maximum being 20 points). These scores were then averaged over the ten sentences leading to a mean maximum possible score of 20 for the test overall. A higher score indicates greater legibility.

Word Spelling

Word spelling ability was measured using the word spelling subtest from the Wide Range Achievement Test IV (WRAT-IV; Wilkinson, & Robertson, 2006). This measure was adapted for group administration, following the published guidelines closely. Children spelled to dictation 13 alphabet letters followed by words graded in difficulty. Children in Year 6 completed all 42 words. Children in all other class year groups completed 36 words. The cut-off for younger children, corresponds to a standard score of 145 for Class Year 5 children in the U.K. It was expected that most children were unlikely to exceed this score. Each correctly spelled word was awarded one point and scoring was discontinued after 10 consecutive errors.The internal reliability of this measure was good (α = .89).

Handwriting Proficiency Screening Questionnaire (HPSQ)

Teacher assessments of handwriting were sought using the Handwriting Proficiency Screening Questionnaire (HPSQ; Rosenblum, 2008). The questionnaire includes 10 questions which measure wellbeing, handwriting fluency, and handwriting legibility. Teachers were asked to observe children writing and to answer 10 questions, such as ‘Is the child’s handwriting unreadable’, using a five-point Likert scale from 0 (*Never*) to 4 (*Always*). A higher score indicates lower handwriting ability. The internal reliability of this measure was good (α = .89).

Procedure

Children completed the measures in specially prepared booklets in typical classroom conditions. The measures were delivered over two sessions not exceeding 60 minutes. All sessions were conducted by a team of at least three trained research assistants who maintained good oversight of children’s work and of their compliance with the instructions.

All children completed all measures. Teacher's ratings on the HPSQ were taken from 107 children. Those scored were selected based on (a) teacher availability to complete the observation and scoring (b) child availability to be observed and scored.

Results

To meet the first aim of this study, to evaluate the reliability and validity of each of the two scales of the SaHLT, each in its own right, we undertook a rigorous psychometric evaluation of the spelling and handwriting legibility subtests. Second, to meet our overarching aim to investigate the concurrent relationship between the spelling and handwriting scales of this assessment tool, we undertook multi-group factor analysis

Evaluation of the Psychometric Properties of the Spelling and Handwriting Legibility Subscales

Developmental Sensitivity

We sought to test the developmental sensitivity of the spelling and handwriting subtests of the test. We did so by comparing the class year (i.e., school grade) means on the total spelling and handwriting scores, respectively (see Table 2 for means and standard deviations for each class year group). As expected, there were moderate-to-large differences in spelling accuracy between class years, $F(3, 1455) = 70.83$, $p < .001$, $\eta_p^2 = .13$. Post-hoc comparisons with Bonferroni corrections revealed spelling accuracy significantly increased with each grade. That is, children in Class Year 6 were more accurate than those in Class Year 5 ($p < .001$), who were more accurate than those in Class Year 4 ($p < .001$), who, in turn, were more accurate than those in Class Year 3 ($p < .001$). There were also moderate-to-large differences in handwriting legibility between class years, $F(3, 1456) = 93.41$, $p < .001$, $\eta_p^2 = .16$. Post-hoc comparisons with Bonferroni corrections showed children in Class Year 6 had significantly better handwriting legibility than children in Class Year 5 ($p < .001$), who had significantly better handwriting legibility than children in Class Year 4 ($p = .021$), who, in turn had significantly more legible writing that children in Class Year 3 ($p < .001$).

[Table 2 here]

Internal Reliability

We assessed internal reliability by examining Cronbach's alpha as a function of each year group and in total for the sentence level scores on both the spelling and total handwriting legibility subtests. The estimates reported in Table 3 show excellent reliabilities for both spelling and handwriting legibility subtests.

[Table 3 here]

Test-retest Reliability

A representative sub-sample of the total number of children in Class Years 3, 4, and 5 completed the SaHLT a second time, approximately seven months after the initial assessment. Responses at the two time points were scored by the same assistant. Test-retest reliability was assessed by calculating the intra-class correlation (ICC) between children's performance on the sentences administered at times one and two. Table 4 reports the ICC and associated 95% confidence intervals for each Year Group on the spelling accuracy (left panels) measures of the SaHLT. As can be seen, the spelling accuracy measure had good to excellent reliability (Koo & Li, 2016). The ICCs of the handwriting subscale demonstrate this measure had moderate-to-good test-retest reliability (Koo & Li, 2016).

[Table 4 here]

Inter-rater Reliability

For spelling, a research assistant and the first author scored a representative sub-sample of the total children tested in Class Years 3 – 6. Inter-rater reliability was evaluated using a two-way random effects ICC. The ICCs and 95% confidence intervals for each year group are reported in Table 5. The ICCs report a very high level of agreement between raters (Koo & Li, 2016). For handwriting, a trained specialist teacher and a research assistant each scored a representative sub-sample of the total number of children tested in Class Years 3 – 6. The sample was selected via stratified (on age and gender) random sampling. Inter-rater reliability was also evaluated using a two-way random effects ICC which are reported in Table 5. The ICCs demonstrate that the SaHLT Handwriting Legibility subscale has a good level of agreement between raters (Koo & Li, 2016).

[Table 5 here]

Convergent Validity

We analysed the convergent validity of the Spelling subtest by comparing it with a well-established measure of single word spelling ability, the WRAT IV Spelling test (Wilkinson & Robertson, 2006). The correlation between the SaHLT Spelling and WRAT IV Spelling was statistically significant ($p < .001$) and strong across all class years (Class Year 3 $r = .76$; Class Year 4 $r = .86$; Class Year 5 $r = .82$; Class Year 6 $r = .77$), with an overall correlation of $r = .84$. This demonstrates good convergent validity for spelling.

We measured the convergent validity of the SaHLT's handwriting legibility measure by examining its association with teacher ratings on the Handwriting Proficiency Screening Questionnaire (HPSQ; Rosenblum, 2008) amongst 107 children (33 children in Class Year 3, 41 children in Class Year 4, and 33 children in Class Year 5) from the sample. Unfortunately, due to logistical issues, we were unable to ascertain teacher responses on the HPSQ for children in Class Year 6. The HPSQ is a valid and reliable screening questionnaire for teachers designed to measure teacher perceptions of a child's handwriting legibility, fluency, and well-being. Well-being in this context is described as including children's willingness to write and capturing fatigues whilst writing (Rosenblum, 2008). In this questionnaire, higher ratings from teachers indicate poorer handwriting. The correlations between the SaHLT Handwriting scores and the HPSQ were statistically significant ($p < .001$) and large in size. The correlations ranged from $r = -.61$ in Class Year 3, $r = -.72$ in Class Year 4, and $r = -.64$ in Class Year 5, giving a total correlation of $r = -.66$ across the sample. The size of these correlations is smaller than those reported for the spelling test, but these correlations were expected as the HPSQ measures aspects of handwriting (e.g., fluency and wellbeing) along with legibility whereas both the spelling measures simply measure spelling accuracy. In sum, these correlations demonstrate good convergent validity of the handwriting subscale.

Factor Structure of Concurrent Spelling and Handwriting Measurement

We sought to examine the factor structure of a concurrent measure of spelling accuracy and handwriting legibility. To do so, we initially ran bivariate correlations between all sub-measures of the SaHLT. We split the spelling measure into three sub-measures: accuracy of function words (n = 22 words; α = .85), accuracy of morphologically simple (uninflected; n = 20 words; α = .93), and accuracy of morphologically complex (inflected; n = 20 words; α = .91) words. For handwriting, this included letter formation (α = .95), letter spacing (α = .94), word spacing (α = .94), and line alignment (α = .95). The correlations reported in Table 6 show strong correlations between the spelling subtest measures. The correlations between the handwriting subtest measures were also strong, but slightly weaker than the spelling measures. The correlations between the spelling and handwriting subtest measures were weaker than the intra-subcomponent measures but were still moderate in size.

[Table 6 here]

We used confirmatory factor analysis (CFA) to examine the factorial structure of the SaHLT (factorial validity) and test whether this structure held over the year groups. The purpose of doing so was to test whether it was possible to measure the two constructs of spelling accuracy and handwriting legibility concurrently in children aged 7–11 years old and then to examine the relationship between the two constructs. In our initial CFA, the model was applied to the whole sample (N = 1,460). The model was run using direct ML estimation in Mplus 8.4 (Muthén & Muthén, 2019). Direct ML was favoured given a small amount (0.78%) of missing data completely at random (MCAR). MCAR status was further confirmed as we found no systematic relationship between the observed and missing cases. Little's MCAR test, implemented using Stata 13.1 (Stata Corp., 2013)yielded χ^2 (4) = 5.13, p = .274. Direct ML is preferred as it uses all the available data (Brown, 2015). Data were approximately normally distributed and there were no significant outliers.

In the model, the three separate sub-measures of spelling: accuracy of function words, accuracy of morphologically simple, and accuracy of morphologically complex words were loaded onto one spelling factor. All the handwriting legibility measures (letter formation, letter spacing, word spacing, and line alignment) were loaded onto a second, handwriting, factor. Furthermore, the residual variances between letter formation and letter spacing were correlated to account for both measures tapping letter level legibility. This model was a reasonable fit to the data, $\chi^2(12) = 130.64$, $p < .001$, RMSEA = .082, SRMR = .031, CFI = .98 and TLI = .96, and all loadings and correlations were significant. Overall, there were high factor loadings of the indicators on their respective factors (> .7), in line with our predictions. The moderate correlations between latent factors suggested no issues with multicollinearity.

The single-group model was then applied to a multigroup CFA (MGCFA), which allows the testing of measurement invariance to assess whether the measures were comparable across the groups. That is, it ensures the model and the indicators were valid measures across all class year groups (Brown, 2015; Milfont & Fischer, 2010). To assess measurement invariance a stepwise procedure of successively more restricted models was used as recommended by Brown (2015), Milfont and Fischer (2010), and Vandenberg and Lance (2000). Using this procedure, the more constrained model is a nested version of the previous model. As such, the new model's goodness-of-fit was examined against the previous, less constrained model. A direct test of fit between the models was completed using a chi-square difference test to ensure the models do not significantly differ from one another.

In addition, a decrease in the CFI magnitude would indicate that the more constrained model be rejected (Cheung & Rensvold, 2002).

Table 7 shows that constraining the factor structure yielded configural invariance equal across groups and produced an acceptable fit. This model acted as the baseline for the further, more constrained models. In the next analysis, the equality of factor loadings (metric invariance) was tested between class years by constraining factor loadings to be equal across groups. Constraining factor loadings gave an overall acceptable model fit. Furthermore, the model did not significantly differ from the configural model, $\chi^2_{diff}(15) = 24.53$, *ns*, nor was there any change in the CFI value. In the next model, the intercepts of the indicators were constrained to be equal across all class years (scalar invariance). This model did not differ from the less constrained metric invariance model as measured by Chi-squared difference, χ^2_{diff} (15) = 0.00, *ns*, nor was there a change in CFI.

Invariance across indicator residuals (differences between measurement error between groups) was not tested as it was deemed to be overly restrictive given there were no theoretical or methodological reasons to expect errors to be equal across class years (see Brown, 2015). Similarly, structural invariance was not tested as performance on indicators did change developmentally (see earlier section on developmental sensitivity). The analyses demonstrate the current solution has measurement invariance, indicating the spelling and handwriting measures were separable and valid measures of their respective constructs across the four class years. The path diagram of the final model the accompanying unstandardized and standardised factor loadings and indicator residual variances for each class year are presented in Figure 1 and Table 8.

All indicators significantly loaded onto their respective factors (see Table 8). Moderate correlations were present between the spelling and handwriting factors for Class Years 3, 4, and 5 and a small correlation for Class Year 6. This further demonstrates the concurrent test was measuring separable, but related constructs. These constructs maintained a similar moderate relationship in Class Years 3-to-5, but the relationship weakened in Class Year 6.

[Table 7 here]

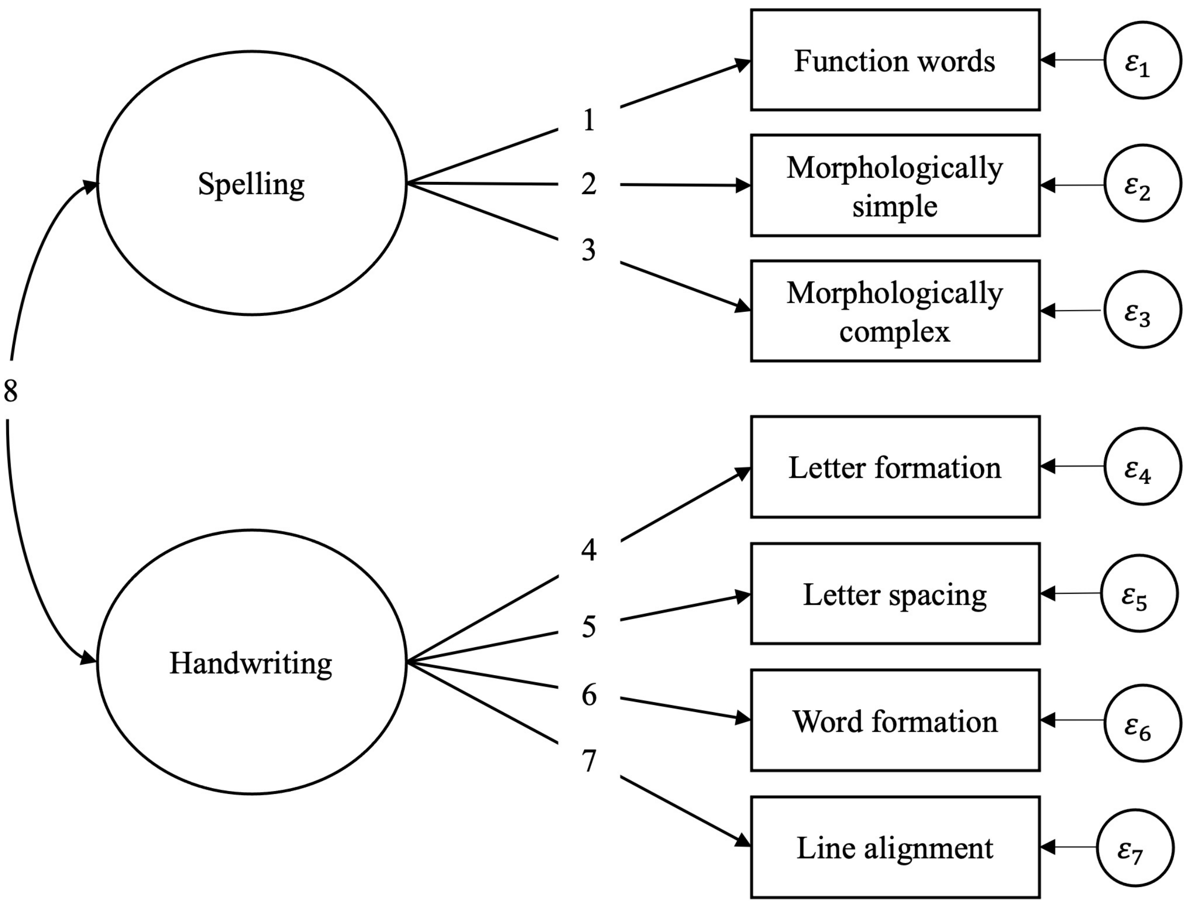

Figure 1 Path model of multi-group confirmatory factor analysis (MGCFA) of the spelling and handwriting subtests of the Spelling and Handwriting legibility Test (*N* = 1460) of four class years (7–11 years old). Path numbers 1–7 correspond to path estimates with residual variances and path number 8 corresponds to the relationship estimates of spelling and handwriting legibility. The corresponding estimates are reported in Table 8

[Table 8 here]

Discussion

We sought to develop and test a measure to investigate the concurrent relationship between spelling and handwriting. First, we examined whether sentence spelling task would be a sensitive, reliable, and valid measure of spelling ability in English. Second, we evaluated whether it was possible to robustly measure spelling and handwriting legibility concurrently. We found that the spelling test was a suitable measure of English spelling, and it was possible to robustly measure spelling and handwriting legibility concurrently in mid-primary-school-aged children. Specifically, the concurrent measure demonstrated sensitivity to age related increases in both spelling accuracy and handwriting legibility. Furthermore, both spelling and handwriting measures had good internal and test-retest reliability, and handwriting had good inter-rater reliability. We also found that the measures of both spelling and handwriting correlated well with respective established tests of spelling and handwriting, confirming convergent validity. In addition, we sought to test the structure and inter-relationships between the constructs. We found the test measured performance on two factors: spelling and handwriting legibility constructs. These were independent of one another but were interrelated. This model applied equally across the four class years, suggesting good stability.

To our knowledge, this is the first assessment of a concurrent measure of sentence spelling and handwriting legibility. We argue that it is theoretically informative, cost and time effective to concurrently measure spelling and handwriting. The SaHLT was able to capture with some sensitivity age related changes in spelling and handwriting legibility. Scores on the sentence spelling test increased as a function of grade, suggesting that the graded increases in the linguistic complexity of the words presented in the sentences adequately captured age-related increases in spelling ability. This supports the view that a sentence spelling task can offer insights about children’s single word spelling ability that are comparable to standard tests of spelling ability in English.

In contrast to spelling, fewer measures of handwriting legibility have been able to detect age related differences in legibility, particularly in older children (e.g., Graham et al., 1998; Gosse et al., 2021). There may be a couple of reasons for this contrasting finding. The most plausible reason for differences in reported developmental trajectories in handwriting is that they capture variations in instructional approaches to handwriting. It follows that in areas where there is a highly consistent instructional approach to handwriting (e.g., Czech Republic; Bartošová et al., 2012) typically developing children are likely to reach their optimum legibility at a younger age. This, in turn, reduces the ability to capture variance in older children. Whereas, in areas where there is less consistency in the instructional approaches to handwriting (e.g., U.S.; Graham et al., 2008), children may not reach their optimum legibility as early and so it easier to capture variance for longer periods of time. Whilst direct cross-education-system studies – which would test this theory – are lacking, indirect evidence is consistent with the above hypothesis. For example, early plateaus of handwriting legibility are reported in systems with a highly structured approach to handwriting (e.g., Belgium; c.f., Gosse et al., 2021) but not in systems with an unstructured approach to handwriting such as the UK and U.S. (Graham et al., 1998). An alternative hypothesis for the between-study differences, which is not mutually exclusive, is that previous methods of estimating handwriting legibility have not had adequate reliability (c.f., Rosenblum et al., 2003). We argue that generalizable handwriting legibility measures should be developed and validated across writing systems to test these theories.

Rosenblum et al. (2003) highlighted several issues regarding the proper assessment of handwriting; these included (a) a lack of clarity in scope of the measures, (b) how practical the measure is to administer, and (c) a lack of robust evaluation of handwriting measures. Unfortunately, since Rosenblum et al. (2003) little progress on dealing with these issues seems to have been made in relation to handwriting legibility. The current study contributes new insights to these issues. The SaHLT can be used for both screening via whole class administration and for individual assessment. We have demonstrated that the test can be used in the classroom and also for research (c.f., Caravolas et al., 2020). Most importantly, though, the current study demonstrates a comprehensive evaluation of the psychometric properties of an analytic handwriting measure.

The current study demonstrates the SaHLT's spelling and handwriting subtests have good psychometric properties. The sentence spelling task has excellent reliability and validity. Performance on the sentence spelling task correlated highly with a well-validated published single word spelling test, suggesting good convergent validity and further support for the utility of a sentence spelling task as an ecologically valid assessment of English spelling ability.

In what may well be the most comprehensive psychometric evaluation of a handwriting legibility measure to date, we found measuring handwriting legibility across four dimensions to be a reliable and valid test of handwriting legibility amongst children aged 7 – 11 years old. The testing and reporting of the reliability and validity of handwriting legibility measures in the literature is highly variable (c.f., Rosenblum et al., 2003). Among the measures which have received some psychometric evaluation, the current four-dimension measure fared well. On inter-rater reliability, the SaHLT handwriting legibility subtest performed similarly to the BHK (Hamstra-Bletz et al, 1987) and better than the SOS (Waevelde et al., 2005) and Ziviani and Elkins (1984). Similarly, on test-retest reliability, the current measure was more reliable than the SOS (Van Waevelde et al., 2005). In addition to reliability, we evaluated several types of validity and found the measure to be valid. Our evaluation of the handwriting subtest's convergent validity found it to be weaker than the BHK's (Hamstra-Bletz et al, 1987). However, it is important to note that we validated our subtest with a measure that tapped both fluency and legibility (the DASH; Barnett et al., 2007) whereas Hamstra-Bletz et al. (1987) correlated the BHK scores with direct teacher ratings of legibility. As such, this discrepancy is likely to stem from the choice of the validation measure used.

We were also able to assess the construct validity of the spelling and handwriting sub-measures using factor analyses. Crucially, these techniques also afforded us the opportunity to directly test the question as to whether it is possible to concurrently measure spelling and handwriting legibility as separable constructs. The factor analysis confirmed the presence of two separable constructs that, moreover, were developmentally stable for 7–11 year olds. Furthermore, the correlations between the constructs revealed that spelling and handwriting legibility are moderately related, and that this relationship is of a similar strength for children aged 7–11 years. However, we did observe a noticeably smaller correlation in the older class year group of children (10–11-year-olds) in this sample. This correlation may have weakened due to the decoupling of spelling and handwriting legibility as the skills become increasingly automatised in older children (Gosse et al., 2021; McCutcheon, 1996). An alternative – but not as compelling – explanation is that the weaker correlation may reflect the smaller *n* (*n* = 82) in this class year.

It may be argued that the observed relationships between spelling and handwriting legibility measures stem from shared method variance, as both measures are derived from the same production. However, in previous research with the SaHLT, we found that a significant relationship obtains between other measures of spelling (e.g., single word spelling) and the SaHLT handwriting legibility dimensions used in the current study (Caravolas et al., 2020). As such, it is unlikely that shared method variance accounts for the relationships found here. The presence of a significant concurrent relationship between spelling and handwriting legibility adds weight to the existing literature also reporting a relationship between these skills in typical and atypical populations (e.g., Caravolas et al., 2020).

There is still more work to be done in understanding the relationship between spelling and handwriting; the measure described and tested here may provide a useful tool for doing so in alphabetic languages. One area for further study is the environmental effects that may influence this relationship, including the curricular emphasis that is placed on handwriting skills in different educational, cultural, and linguistic systems. Such direct comparisons would allow us to further understand the language-, orthography- and culture-general relationships between spelling and handwriting, and the between-study (and country) differences in the timing of plateauing of handwriting legibility, which seems to occur earlierin some areas (e.g., Belgium) than in others (e.g., U.K.). The current test could be readily adapted to other languages; for example, a parallel spelling measure has already been created in Czech (Caravolas et al., 2005). Furthermore, the current handwriting legibility dimensions – created for assessing legibility in a highly variable educational context – could well be a suitable measure for handwriting legibility components which are universal across languages and education-systems.

Whilst this study has demonstrated that it is possible to concurrently measure spelling and handwriting legibility and their associations, an important next step is to establish causality in their relationship (see Hulme & Snowling, 2012). That is, the findings presented here do not establish causality or the direction of causality. To this aim, further longitudinal and intervention studies are needed (see Gosse et al., 2021; Pritchard et al., 2021).

In sum, we sought to investigate the concurrent relationship between handwriting and spelling. We found that a sentence spelling task with stimuli varying in their linguistic complexity is sensitive to grade/age-related differences, and offers a reliable and valid alternative to single word spelling, the typical assessment method in English. Furthermore, our study demonstrated that it is possible to robustly measure spelling and handwriting legibility concurrently, and, that they are separable, but related abilities. The measure we evaluated in this paper offers a cost and time saving measure of both constructs simultaneously. Furthermore, this soon-to-be freely available measure (for further information please see: http://www.eldel-bll.uk/sahlt/) can be used in future investigations of the relationship between spelling and handwriting legibility.

References

Abbott, R. D., & Berninger, V. W. (1993). Structural equation modeling of relationships among developmental skills and writing skills in primary- and intermediate-grade writers. *Journal of Educational Psychology*, *85*(3), 478–508. https://doi.org/10.1037//0022-0663.85.3.478

Abbott, R. D., Berninger, V. W., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. *Journal of Educational Psychology*, *102*(2), 281–298. https://doi.org/10.1037/a0019318.

Adi-Japha, E., Landau, Y. E., Frenkel, L., Teicher, M., Gross-tsur, V., & Shalev, R. S. (2007). ADHD and dysgraphia: Underlying mechanisms. *Cortex*, *43*(6), 700–709. https://doi.org/10.1016/S0010-9452(08)70499-4.

Afonso, O., Suárez-Coalla, P., González-Martín, N., & Cuetos, F. (2018). The impact of word frequency on peripheral processes during handwriting: A matter of age. *Quarterly Journal of Experimental Psychology*, *71*(3), 695–703. https://doi.org/10.1080/17470218.2016.1275713

Asher, A. V. (2006). Handwriting instruction in elementary schools. *The American Journal of Occupational Therapy*, *60*(4), 461-471. https://doi.org/10.5014/ajot.60.4.461

Alamargot, D., & Morin, M. F. (2015). Does handwriting on a tablet screen affect students' graphomotor execution? A comparison between grades two and nine. *Human Movement Science*, *44*(1), 32–41. https://doi.org/10.1016/j.humov.2015.08.011

Barnett, A., Henderson, S. E., Scheib, B., & Schulz, J. (2007). The detailed assessment of speed of handwriting (DASH) [Assessment instrument]. London, UK: Pearson.

Barnett, A. L., Prunty, M., & Rosenblum, S. (2018). Development of the Handwriting Legibility Scale (HLS): A preliminary examination of reliability and validity. *Research in Developmental Disabilities*, *72*, 240–247. https://doi.org/10.1016/j.ridd.2017.11.013

Bartošová, I., Manĕnová, M., & Třečková, E. (2012). The new Comenia script to schools. *Procedia-Social and Behavioral Sciences*, *69*, 2228–2236. https://doi.org/10.1016/j.sbspro.2012.12.192

Berninger, V. W., & Winn, W. D. (2006). Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 96–114). The Guilford Press.

Bosse, M. L., Brissaud, C., & Le Levier, H. (2021). French pupils' lexical and grammatical spelling from sixth to ninth grade: A longitudinal study. *Language and Speech*, *64*(1), 224-249. https://doi.org/10.1177/0023830920935558

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. The Guilford Press.

Calhoon, M. B., & Masterson, J. J. (2011). Lexical analysis of words on commonly used standardized spelling assessments. *Assessment for Effective Intervention*, *36*(2), 80–93. https://doi.org/10.1177/1534508410380136

Caravolas, M. (2022). Reading and reading disorder in alphabetic languages. In M. Snowling, C. Hulme, & K. Nation (Eds.). *The Science of Reading: A Handbook* (2nd Edition) (pp. 337-353). John Wiley and Sons Ltd.

Caravolas, M. (2004). Spelling development in alphabetic writing systems: A cross-linguistic perspective. *European Psychologist*, *9*(1), 3–14. https://doi.org/10.1027/1016-9040.9.1.3

Caravolas, M. & Kessler, B. (2016, July 13–16). *Spelling development in alphabetic orthographies of varying consistency: measurement matters* [Poster presentation]. Twenty Third Annual Meeting of the Society for the Scientific Studies of Reading, Porto, Portugal. http://spell.psychology.wustl.edu/CaravolasSSSR2016/CaravolasSSSR2016.pdf

Caravolas, M., & Volín, J. (2001). Phonological spelling errors among dyslexic children learning a transparent orthography: The case of Czech. *Dyslexia*, *7*(4), 229–245. https://doi.org/10.1002/dys.206

Caravolas, M., Volín, J., & Hulme, C. (2005). Phoneme awareness is a key component of alphabetic literacy skills in consistent and inconsistent orthographies: Evidence from Czech and English children. *Journal of Experimental Child Psychology*, *92*(2), 107–139. https://doi.org/10.1016/j.jecp.2005.04.003

Caravolas, M., Downing, C., Hadden, C. L., & Wynne, C. (2020). Handwriting legibility and its relationship to spelling ability and age: evidence from monolingual and bilingual children. *Frontiers in Psychology*, 11, 1097. https://doi.org/10.3389/fpsyg.2020.01097

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*(4), 751–774. https://doi.org/10.1006/jmla.2000.2785

Caravolas, M., Lervåg, A., Mousikou, P., Efrim, C., Litavsky, M., Onochie-Quintanilla, E., … Hulme, C. (2012). Common patterns of prediction of literacy development in different alphabetic orthographies. *Psychological Science*, *23*(6), 678–86. https://doi.org/10.1177/0956797611434536

Caravolas, M., Mikulajová, M., Defior, S., & Seidlová Málková, G. (2018). *Tests*. *Multilanguage Assessment Battery of Early Literacy*. MABEL. https://www.eldel-mabel.net/test/

Caravolas, M., Mikulajová, M., Vencelová, L. (2008). Súbor testov na *hodnotenie pravopisu pre zakladne skoly a klinicku prax.* [Assessment battery of orthographic skills for primary schools and clinical practice.] Bratislava: SAL, s.r.o.

Cassar, M., & Treiman, R. (1997). The beginnings of orthographic knowledge: Children's knowledge of double letters in words. *Journal of Educational Psychology*, *89*(4), 631.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902

Daffern, T., & Ramful, A. (2020). Measurement of spelling ability: construction and validation of a phonological, orthographic and morphological pseudo-word instrument for students in Grades 3–6. *Reading and Writing*, *33*(3), 571–603. https://doi.org/10.1007/s11145-019-09976-1

Department for Education (2014). *National curriculum in England: framework for key stages 1 to 4.* UK Government. https://www.gov.uk/government/publications/national-curriculum-in-england-framework-for-key-stages-1-to-4/the-national-curriculum-in-england-framework-for-key-stages-1-to-4

Downing, C. (2018). *Understanding writing difficulties amongst children with neurodevelopmental disorders: The cases of dyslexia and/or developmental coordination disorder (DCD)* (Doctoral dissertation, Bangor University). Retrieved from https://research.bangor.ac.uk/portal/en/theses/understanding-writing-difficulties-amongst-children-with-neurodevelopmental-disorders(24b0633f-1586-4ac0-b758-89a7b1e0ba8a).html

Elliot, C. (2011). British ability scales: Third edition (BAS-III) [Assessment instrument]. London, UK; GL Assessment.

Feder, K. P., & Majnemer, A. (2007). Handwriting development, competency, and intervention. *Developmental Medicine and Child Neurology*, *49*, 312–317.

Gosse, C., Parmentier, M., & Van Reybroeck, M. (2021). How Do Spelling, Handwriting Speed, and Handwriting Quality Develop During Primary School? Cross-Classified Growth Curve Analysis of Children's Writing Development. *Frontiers in Psychology*, 2927.

Graham, S., Berninger, V. W., Weintraub, N., & Schafer, W. (1998). Development of handwriting speed and legibility in grades 1–9. *The Journal of Educational Research*, *92*, 42–52. https://doi.org/10.1080/00220679809597574.

Graham, S., Harris, K. R., Mason, L., Fink-Chorzempa, B., Moran, S., & Saddler, B. (2008). How do primary grade teachers teach handwriting? A national survey. *Reading and Writing*, *21*(1–2), 49–69. https://doi.org/10.1007/s11145-007-9064-z

Graham, S., Struck, M., Santoro, J., & Berninger, V. W. (2006). Dimensions of good and poor handwriting legibility in first and second graders: Motor programes, visual-spatial arrangement, and letter formation parameter setting. *Developmental Neuropsychology*, *29*(1), 43–60. https://doi.org/10.1207/s15326942dn2901_4

Hamstra-Bletz, E., de Bie, J., & den Brinker, B. P. L. M. (1987). Beknopte beoordelingsmethodevoor kinderhandschriften: Experimentele versie [The concise assessment scale for children's handwriting: Experimental version]. Lisse: Swets & Zeitlinger.

Hamstra-Bletz, L., & Blöte, A. W. (1990). Development of handwriting in primary school: A longitudinal study. *Perceptual and motor skills*, *70*(3), 759–770. https://doi.org/10.2466/pms.1990.70.3.759

Hulme, C., & Snowling, M. J. (2012). *Developmental disorders of language learning and cognition*. John Wiley & Sons.

Kandel, S., & Perret, C. (2015). How does the interaction between spelling and motor processes build up during writing acquisition? *Cognition*, *136*, 325–336. https://doi.org/10.1016/j.cognition.2014.11.014

Kandel, S., Lassus-sangosse, D., Grosjacques, G., & Perret, C. (2017). The impact of developmental dyslexia and dysgraphia on movement production during word writing. *Cognitive Neuropsychology*, *34*(3–4), 219–251.

Kapoor, A. K., & Saini, M. (2017). Handwriting as a means of cultural identity. *Journal of Forensic Science and Criminal Investigation, 3*(1). https://doi.org/10.19080/JFSCI.2017.03.555605

Kessler, B., & Treiman, R. (2003). Is English spelling chaotic? Misconceptions concerning its irregularity. *Reading Psychology*, *24*(3–4), 267–289. https://doi.org/10.1080/02702710390227228

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. http://dx.doi.org/10.1016/j.jcm.2016.02.012

Kučerová, O., & Kucharská, A. (2018). Acquiring penmanship and writing skills from the first to fifth grade of primary school: Joined-up writing vs. Comenia Script. *Journal of Language and Cultural Education*, *6*(2), 1–13. https://doi.org/10.2478/jolace-2018-0012

Lambert, E., & Quémart, P. (2019). Introduction to the special issue on the dynamics of written word production: methods, models, and processing units. *Reading and Writing*, *32*(1), 1–12. https://doi.org/10.1007/s11145-018-9929-3

Masterson, J. J., & Apel, K. (2010). The spelling sensitivity score: Noting developmental changes in spelling knowledge. *Assessment for Effective Intervention*, *36*(1), 35–45. https://doi.org/10.1177/1534508410380039

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review*, *8*(3), 299–325. https://doi.org/10.1007/BF01464076

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of psychological research*, *3*(1), 111–130. https://doi.org/10.21500/20112084.857

Muthén, L. K.., & Muthén, B. O. (2019). Mplus User's Guide (Eigth Edition). Los Angeles, CA: Muthén & Muthén.

Nunes, T., & Bryant, P. (2009). *Children's reading and spelling: Beyond the first steps*. John Wiley & Sons.

Nunes, T., Bryant, P., & Bindman, M. (1997). Morphological spelling strategies: Developmental stages and processes. *Developmental Psychology*, *33*(4), 637–649. https://doi.org/10.1037/0012-1649.33.4.637

Olive, T. (2014). Toward a parallel and cascading model of the writing system: A review of research on writing processes coordination. *Journal of Writing Research*, *6*(2), 173–194. https://doi.org/10.17239/jowr-2014.06.02.4

Pontart, V., Bidet-Ildei, C., Lambert, E., Morisset, P., Flouret, L., & Alamargot, D. (2013). Influence of handwriting skills during spelling in primary and lower secondary grades. *Frontiers in Psychology*, *4*, 1–9. https://doi.org/10.3389/fpsyg.2013.00818

Pritchard, V. E., Malone, S. A., & Hulme, C. (2021). Early handwriting ability predicts the growth of children's spelling, but not reading, skills. *Scientific Studies of Reading*, *25*(4), 304–318. https://doi.org/10.1080/10888438.2020.1778705.

Prunty, M., & Barnett, A. L. (2017). Understanding handwriting difficulties: A comparison of children with and without motor impairment. *Cognitive Neuropsychology*, *34*(3–4), 205–218. https://doi.org/10.1080/02643294.2017.1376630

Quémart, P., & Lambert, E. (2019). The influence of the morphological structure of words on the dynamics of handwriting in adults and fourth and sixth grade children. *Reading and Writing*, *32*(1), 175–195. https://doi.org/10.1007/s11145-017-9762-0

Rosenblum, S. (2008). Development, reliability, and validity of the Handwriting Proficiency Screening Questionnaire (HPSQ). *The American Journal of Occupational Therapy*, *62*(3), 298–307. https://doi.org/10.5014/ajot.62.3.298

Rosenblum, S., & Livneh-Zirinski, M. (2008). Handwriting process and product characteristics of children diagnosed with developmental coordination disorder. *Human* *movement science*, *27*(2), 200–214. https://doi.org/10.1016/j.humov.2008.02.011

Rosenblum, S., Weiss, P. L., & Parush, S. (2003). Product and process evaluation of handwriting difficulties: A review. *Educational Psychology Review*, *15*(1), 41–81.

Schrank, F. A., & Wendling, B. J. (2018). The Woodcock-Johnson IV [Assessment instrument].

Smits-Engelsman, B. C., Niemeijer, A. S., & van Galen, G. P. (2001). Fine motor deficiencies in children diagnosed as DCD based on poor grapho-motor ability. *Human Movement Science*, *20*(1–2), 161–182. https://doi.org/10.1016/S0167-9457(01)00033-1

Smits-Engelsman, B. C., Stevens, M., Vrenken, I., & Van Hagen, A. (2005). Systematische Opsporing Schrijfproblemen (SOS): Een hulpmiddel voor leerkrachten bij het signaleren van motorische schrijfproblemen van leerlingen in het Basis en Speciaal Onderwijs.[Systematic screening of handwriting problems (SOS): An instrument for teachers for screening of handwriting problem of children in primary school and special education]. *Kinderfysiotherapie*, *17*, 16–21.

Stainthorp, R., Barnett, A., Henderson, S., & Scheib, B. (2006). *Handwriting policy and practice in English primary schools: An exploratory study (issues in practice)*. Institute of Education.

StataCorp. 2013. *Stata Statistical Software* (Version 13.1) [Computer software]. StataCorp LP. https://www.stata.com/

Suárez-Coalla, P., Afonso, O., Martínez-García, C., & Cuetos, F. (2020). Dynamics of Sentence Handwriting in Dyslexia: The Impact of Frequency and Consistency. *Frontiers in Psychology*, *11*, 319. https://doi.org/10.3389/fpsyg.2020.00319.

Sumner, E., Connelly, V., & Barnett, A. L. (2014). The influence of spelling ability on handwriting production: Children with and without dyslexia. *Journal of Experimental*

Psychology. Learning, Memory, and Cognition, *40*(5), 1441–7. https://doi.org/.1037/a0035785

Thibon, L. S., Gerber, S., & Kandel, S. (2018). The elaboration of motor programs for the automation of letter production. *Acta Psychologica*, *182*, 200–211. https://doi.org/10.1016/j.actpsy.2017.12.001

Treiman, R. (1993). *Beginning to spell: A study of first-grade children*. Oxford University Press

Treiman, R. (2017). Learning to spell: Phonology and beyond. *Cognitive Neuropsychology*, *34*(3–4), 83–93. https://doi.org/10.1080/02643294.2017.1337630

Treiman, R., & Bourassa, D. C. (2000). The development of spelling skill. *Topics in language disorders*, *20*(3), 1–18.

Treiman, R., Cassar, M., & Zukowski, A. (1994). What types of linguistic information do children use in spelling? The case of flaps. *Child Development*, 65(5), 1318–1337. https://doi.org/10.1111/j.1467-8624.1994.tb00819.x

Treiman, R., & Kessler, B. (2014). *How children learn to write words*. Oxford University Press.

Treiman, R., Kessler, B., & Caravolas, M. (2019). What methods of scoring young children's spelling best predict later spelling performance? *Journal of Research in Reading*, *42*(1), 80–96. https://doi.org/10.1111/1467-9817.12241

van Galen, G. P. (1991). Handwriting: psychomotor Issues for a theory. *Human Movement Science*, *10*, 165–191.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions practices and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70.

Van Waelvelde, H., Hellinckx, T., Peersman, W., & Smits-Engelsman, B. C. (2012). SOS: a screening instrument to identify children with handwriting impairments. *Physical and Occupational Therapy in Pediatrics*, *32*(3), 306–319. https://doi.org/10.3109/01942638.2012.678971

Wilkinson, G. S., & Robertson, G. J. (2006). The Wide Range Achievement Test-4 (WRAT-4) [Assessment instrument]. Lutz, FL: Psychological Assessment Resources.

Ziviani, J., & Elkins, J. (1984). An evaluation of handwriting performance. *Educational review*, *36*(3), 249–261. https://doi.org/10.1080/0013191840360304

Table 1

Description of the Evaluation Sample

Class year	*N*	*n* by gender		Age (months)	
		Boys	Girls	Mean	S.D.
3	494	237	257	99.12	5.1
4	435	222	213	110.11	3.98
5	450	248	202	122.36	3.73
6	82	41	41	129.62	3.69

Table 2

Descriptive Statistics of the Spelling Accuracy and Handwriting Legibility Measures as a Function of Class Year

		n	*M*	*SD*	Range	Max score
			Spelling accuracy			
Class year						
	3	492	38.04	12.38	3 – 62	
	4	435	42.57	10.13	8 – 61	
	5	450	46.46	9.99	8 – 62	62
	6	82	52.29	6.98	23 – 61	
Overall		1459	41.64	12.53	0 – 62	
			Handwriting legibility			
Class year						
	3	494	12.40	2.00	6.2 – 17.9	
	4	435	13.02	1.95	6.7 – 19.7	
	5	450	13.42	2.03	6.0 – 19.6	20
	6	82	16.35	2.54	9.6 – 20	
Overall		1461	13.04	2.37	4 – 20	

Note. Max score is the maximum achievable score.

Table 3

Cronbach's Alpha Estimates for the Spelling and Handwriting Sub-measures as a Function of Class Year

		Spelling	Handwriting
Class year			
	3	.95	.97
	4	.95	.97
	5	.96	.98
	6	.92	.98
Overall		.96	.98

Table 4

Test-retest Reliabilities of the Spelling Measures for Class Years Three, Four, and Five

	Spelling accuracy				Handwriting legibility			
	Delay[a]	*n*	ICC	95% CI	Delay[a]	*n*	ICC	95% CI
Class year								
3	7.5	18	.87	[.65, .95]	7.6	10	.68	[-.26, .92]
4	7.7	24	.91	[.78, .95]	7.1	19	.76	[.39, .91]
5	6.9	23	.91	[.79, .96]	7.6	14	78	[.33, .93]
Overall	7.4	64	.90	[.84, .94]	7.4	43	.76	[.57, .87]

Note. [a]Average number of months between time 1 and time 2. ICC = Two-way mixed effects intra-class correlation. CI = Confidence interval.

Table 5

Inter-rater Reliabilities of the Handwriting Legibility Measure as a Function of Class Year

		Spelling		Handwriting	
	n	ICC	95% CI	ICC	95% CI
Class year					
3	13	.99	[.99, .99]	.84	[.47, .95]
4	19	1.00	[0, 0]	.86	[.63, .95]
5	15	.99	[.99, .99]	.77	[.31, .92]
6	12	1.00	[0, 0]	.79	[.35, .93]
Overall	71	.99	[.99, .99]	.81	[.69, .90]

Note. ICC = Two-way random effects intra-class correlation. CI = Confidence interval.

Table 6

Bivariate Correlations of the Spelling Accuracy and Handwriting Legibility Measures Across the Sample

	1	2	3	4	5	6	7	8
Spelling accuracy								
1. Function								
2. Morphologically simple	.69***							
3. Morphologically complex	.66***	.89***						
4. Total spelling	.63***	.89***	.89***					
Handwriting legibility								
5. Letter formation	.29***	.42***	.45***	.45***				
6. Letter spacing	.25***	.31***	.32***	.34***	.71***			
7. Word spacing	.32***	.36***	.37***	.40***	.57***	.63***		
8. Line alignment	.28***	.31***	.32***	.34***	.62***	.57***	.58***	
9. Total legibility	.34***	.41***	.43***	.46***	.86***	.86***	.82***	.83***

Note. ***$p < .001$.

Table 7

Goodness-of-Fit Estimates for Single- and Multi-Group Confirmatory Factor Analysis Models of Measures Loading Spelling and Handwriting

	χ^2	*df*	RMSEA	SRMR	CFI	TLI	χ^2_{diff}	Δ_{df}
Single-group solution	130.64***	12	.082	.031	.98	.96	-	-
Multi-group solutions								
Full configural invariance	190.21***	48	.090	.038	.97	.95	-	-
Full metric invariance	214.74**	63	.081	.043	.97	.96	24.53	15
Full scalar invariance	214.74***	78	.069	.043	.97	.97	0.00	15

Note. *N* = 1460. χ^2_{diff} = nested difference between the restricted solution and the preceding less-restricted solution. RMSEA = root mean square of error approximation. 90% CI = 90% confidence intervals for RMSEA. SRMR = standardised root mean square residual. CFI = comparative fit index. TLI = Tucker-Lewis index.
**p* < .05, ** *p* < .01, ****p* < .001.

Table 8

Unstandardised and Standardised Factor Loadings of Each Class Year in the Multi-Group Factor Analysis

	Unstandardised		Standardised							
			Class Year 3		Class Year 4		Class Year 5		Class Year 6	
Path	Estimate	Residual	Estimate	Residual	Estimate	Residual	Estimate	Residual	Estimate	Residual
Spelling										
1 Function	1.00(.00)	.60(.10)	.72(.02)	.48(.03)	.70(.02)	.51(.03)	.71(.02)	.51(.03)	.65(.05)	.58(.06)
2 Morph. simple	1.35(.04)	.20(.07)	.98(.01)	.04(.02)	.92(.01)	.16(.02)	.94(.01)	.17(.02)	.89(.04)	.21(.06)
3 Morph. complex	1.29(.04)	.28(.07)	.92(.01)	.16(.02)	.91(.01)	.17(.02)	.92(.01)	.16(.02)	.85(.04)	.28(.06)
Handwriting										
4 Letter formation	1.00(.00)	.43(.10)	.74(.03)	.45(.04)	.79(.03)	.38(.04)	.80(.03)	.36(.04)	.80(.05)	.36(.07)
5 Letter spacing	.95(.04)	.20(.07)	.70(.03)	.51(.04)	.73(.03)	.46(.04)	.74(.03)	.45(.04)	.88(.04)	.23(.07)
6 Word spacing	.86(.05)	.35(.07)	.66(.03)	.56(.04)	.66(.03)	.57(.04)	.67(.03)	.55(.04)	.79(.04)	.38(.07)
7 Line alignment	.90(.04)	.51(.10)	.69(.03)	.52(.04)	.70(.03)	.51(.04)	.70(.03)	.51(.04)	.74(.04)	.45(.07)
Spelling and handwriting	.13(.07)	-	.41(.05)	-	.42(.05)	-	.40(.05)	-	.24(.12)	-

Note. Path numbers correspond to those presented in the path diagram (Figure 1). Morph. = morphologically Residuals correspond to the standardised indicator residual variances. Standard errors are reported in parentheses next to the loading/residual. All factor loadings $p < .001$.