

Evaluating the Spelling and Handwriting Legibility Test (SaHLT):

Downing, Cameron; Caravolas, Marketa

Reading and Writing

DOI: 10.1007/s11145-022-10402-2

E-pub ahead of print: 01/03/2023

Peer reviewed version

Cyswllt i'r cyhoeddiad / Link to publication

Dyfyniad o'r fersiwn a gyhoeddwyd / Citation for published version (APA): Downing, C., & Caravolas, M. (2023). Evaluating the Spelling and Handwriting Legibility Test (SaHLT): a tool for the concurrent assessment of spelling and handwriting. . *Reading and Writing*. Advance online publication. https://doi.org/10.1007/s11145-022-10402-2

Hawliau Cyffredinol / General rights Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

· Users may download and print one copy of any publication from the public portal for the purpose of private study or research.

- You may not further distribute the material or use it for any profit-making activity or commercial gain
 You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

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)

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

Corresponding author:

Dr Cameron Downing (<u>c.downing@leedstrinity.ac.uk</u>) 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 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].

1 The abilities to spell and to write legibly are foundational literacy skills. They set the 2 stage for the development of the complex, higher-level skills of text writing, so critical for 3 educational success (Berninger & Winn, 2006; Olive, 2014). Impairments in one or both 4 skills are related to broader writing difficulties in children with developmental disorders such 5 as dyslexia (e.g., Sumner et al., 2014), developmental coordination disorder (e.g., Prunty & 6 Barnett, 2017), and attention deficit hyperactivity disorder (e.g., Adi-Japha et al., 2007). 7 While spelling and handwriting are under the control of different domains (i.e., spelling -8 primarily cognitive; handwriting - primarily perceptuomotor), emerging research evidence 9 suggests that they are inter-related in typically developing children (e.g., Afonso et al., 2018; 10 Caravolas et al., 2020; Kandel et al., 2017) and in children with developmental disorders such 11 as dyslexia (e.g., Kandel et al., 2017; Suárez-Coalla et al., 2020; Sumner et al., 2014). Yet, 12 the two skills are typically assessed by separate measures that are designed to focus only on 13 each skill in isolation. This practice makes it more difficult, time consuming, and costly to 14 gauge the spelling and handwriting profiles within individuals and to understand the nature of 15 their potential difficulties across the two domains. In the present paper, we report a study that 16 concurrently investigated the development of spelling ability and handwriting legibility, and 17 the nature of the relationship between them, in primary-school-aged children using a single 18 measurement tool, the Spelling and Handwriting Legibility Test (SaHLT). In what follows, 19 we summarise what is known about the development of spelling and handwriting ability, and 20 then briefly review existing approaches to assessing spelling accuracy and handwriting 21 legibility before turning to the development of the SaHLT.

22 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).

8 The English orthography, which is the focus of the present study, is 9 phonographemically less consistent than most other alphabetic orthographies (Caravolas, 10 2022; Caravolas & Kessler, 2016). Many children master this skill of *phonological recoding* 11 by the second year of schooling, but measurable individual differences in phoneme-grapheme 12 encoding skills persist among English-speaking children beyond this point (Caravolas et al., 13 2001; Caravolas & Kessler, 2016). Children with spelling difficulties may continue to 14 struggle to represent attributes of written words that are more complex, such as consonant 15 clusters, complex vowels (diphthongs), unstressed syllables, and polysyllabic words (e.g., 16 Treiman, 2017). For this reason, tests assessing children's skills often contain words varying in terms of such attributes (Caravolas, 2004; Treiman, 1993). 17

18 From a psycholinguistic perspective, the inconsistent and complex nature of the 19 English orthography requires that learners have not only good phonological skills, but also 20 some knowledge about the morpho-phonological basis of the system (e.g., Nunes & Bryant, 21 2009). Moreover, learners need to be sensitive to the distributional properties of the letters 22 and letter patterns occurring in printed words, many of which have no clear connection to 23 phonology or morphology, but instead reflect archaic or foreign pronunciations (Treiman & 24 Kessler, 2014). Therefore, conventional spelling competence requires the integration of higher-order knowledge about the morphological (Nunes et al., 1997; Treiman et al., 1994) 25

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.

7 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 8 9 obtained from samples of children's free writing, and by the recognition of correctly spelled 10 words from incorrectly spelled alternatives. The first of these methods is arguably the most 11 frequently used, with numerous reliable, standardised, single-word spelling tests being 12 available for clinical and educational use (see Calhoon & Masterson, 2011). Typically, the 13 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 14 15 (canonical) sound-letter mappings that can be derived by "sounding out" the words' 16 phonemes. Measures are usually scored for overall accuracy, and as such they tend to be 17 highly reliable and sensitive to age/schooling levels but provide limited insights into the 18 causes of children's misspellings. More recent approaches (e.g., Caravolas et al., 2018; 19 Daffern & Ramful, 2020; Masterson & Apel, 2010; Treiman & Bourassa, 2000) select test 20 items that reflect the variety of linguistic and orthographic sources of sound-letter 21 inconsistency (e.g., graphotactic, morphophonological, etymological). This approach is 22 traditional in many languages with more consistent orthographies (e.g., Czech: Caravolas & 23 Volín, 2001, 2005; Slovak: Caravolas et al., 2008; Greek: Sideridis et al., 2008) and it 24 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, 25

5

where words graded in difficulty were embedded in dictated sentences, was also adopted in
 the SaHLT tool we describe here.

3 While the optimal method of spelling assessment is likely to depend on the aims of 4 the test and the context of the learner, the single-word format is adequate for estimating 5 general spelling skills in English, assuming well-selected word lists, because the language 6 has relatively low morphological (especially inflectional) complexity, as compared to many 7 languages. Consequently, individual word spellings are not so affected by grammatical 8 agreement among neighbouring words in phrases and sentences. In contrast, in many 9 languages with alphabetic orthographies and with rich and more complex morphology and 10 grammar, word spellings may alter depending on the syntactic/grammatical context of 11 neighbouring words, and accordingly, spelling tests often comprise short written texts, or 12 phrases containing the target words and their disambiguating context – which itself may be 13 phonographemically inconsistent (e.g., Bosse et al., 2021; Caravolas et al., 2008). Nevertheless, the encoding of morphological information in word spellings and the existence 14 15 of heterographic homophones in the English orthography typically require aurally presented 16 carrier sentences to provide context in single word spelling tests. Arguably, tests requiring the 17 writing of full sentences better approximate a typical writing situation and, as such, provide 18 an ecologically valid assessment of spelling (Department for Education, 2014). However, 19 sentence writing methods for assessing English spelling are seldom used in standardised tests 20 while single word dictation is the primary assessment method (Treiman et al., 2019). 21 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 22

comparison of English and the highly inflected Czech language (Caravolas & Volín, 2001;

24 Caravolas et al., 2005). It adopts the idea of graded word lists of the English tradition and

25 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.

6 Handwriting Legibility Development and its Assessment

7 Handwriting ability is typically considered in terms of fluency and legibility. 8 Handwriting fluency is the ability to produce written script at speed without undue hesitation 9 or effort. Legibility pertains to the ability to produce written scripts such that they are easily 10 decipherable to the reader; albeit norms for well-formed, legible scripts are to some extent 11 culturally determined (Kapoor & Saini, 2017). Whereas handwriting legibility is difficult to 12 assess objectively, handwriting fluency is relatively easy to measure by recording the number 13 of units (e.g., letters or words) written within a specified time frame (e.g., Alamargot & 14 Morin, 2015; Pontart et al., 2013). The relative ease in measuring handwriting fluency using 15 such tools as the Detailed Assessment of Speed of Handwriting (Barnett et al., 2007) has led 16 to it being more extensively studied and tested in the classroom and clinic (Lambert & 17 Quémart, 2019). However, fluent handwriting requires sufficiently consolidated 18 representations of letters and words (Thibon et al., 2018) and so fluency measures may not be 19 reliable for less able (younger and impaired) learners who have yet to consolidate 20 representations. In contrast, handwriting legibility reflects the quality of these 21 representations, at least in part (e.g., Pritchard et al., 2021), and so may be more reliable and 22 appropriate for less skilled writers. Handwriting fluency and legibility are separable skills 23 with different developmental trajectories (Authors, submitted, Gosse et al., 2021; Graham et 24 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, 25

2014). It is therefore important to measure handwriting legibility in addition to handwriting
 fluency.

3 Relatively fewer investigations have focused on handwriting legibility than on 4 handwriting fluency. Legibility measures can be categorized as global or analytic. Global 5 measures are those where criteria are applied to the written product (e.g., a multi-sentence 6 text) as a whole, whereas analytic measures typically apply criteria to specific attributes of 7 legibility, such as formation, size, slanting, spacing, and straightness of letters, but different 8 scales operationalise these attributes differently. In their review, Rosenblum et al. (2003) 9 have argued that analytic measures have become more popular than global measures 10 (although, see Barnett et al., 2018 for a recent example of a global handwriting measure). 11 Despite many measures broadly agreeing on the attributes of legibility, the theoretical 12 rationale for selecting these attributes remains unclear. Furthermore, Rosenblum et al. (2003) 13 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 14 15 could be used, be it screening and assessment of handwriting difficulties and/or tracking 16 typical development), (b) practical administration of the measure (i.e., the ease of use in 17 classroom, clinic, and research settings), and (c) how well the measure has been evaluated in 18 terms of its psychometric properties. Indeed, of the 13 analytic measures reviewed, only 2 19 scales had undergone a comprehensive evaluation of their reliability, validity, and 20 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

1 that their scale was sensitive to age-related changes in legibility and undertook efforts to 2 evaluate the reliability and validity of the measure. Whilst inter-rater reliability of nonlinguistic (letter-like) symbols (r = .88 - .98) fell above the authors' own desirable cut-off (r3 4 = .8), the inter-rater reliability of linguistic items (letters and words) – and arguably most 5 ecologically valid measures of legibility – fell below the threshold (.69-.76). Similarly, test-6 retest reliability (measured one week apart) overall was below the authors' threshold (.62-7 .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 = .46-8 9 .76), but this method does not demonstrate criterion validity, rather, it further demonstrates 10 the reliability of the measure. Finally, principal component analysis did verify that the criteria 11 measured different components of handwriting legibility and the scale was sensitive to 12 developmental changes. To our knowledge, there has been no further work to improve the 13 scale's psychometric properties.

14 A second well-evaluated scale is a Dutch language assessment of handwriting: 15 Beknopte Beoordelingsmethode voor Kinder Handschriften (BHK; Concise Assessment 16 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 17 18 copy a passage for 5 minutes, which is then scored using 13 criteria which are operationalised 19 through a mixture of Likert type and yes/no responses. This measure has been well evaluated 20 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) 21 22 and internal consistency, although relatively low ($\alpha = .52$), to be acceptable (Hamstra-Bletz 23 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 24 25 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.

3 To address the practical issues of the BHK, Smits-Engelsman et al. (2005) developed 4 the Systematische Opsporing van Schrijfmotorische problemen (SOS) by identifying and 5 reformulating the six most discriminating items from the BHK. The SOS has been found to 6 be developmentally sensitive for children aged 7-to-11 years, with excellent interrater 7 reliability (ICC = .77) and moderate test-retest reliability (ICC = .69; Van Waelvelde et al., 8 2012). SOS scores correlate highly with BHK scores (r = .70) suggesting good convergent 9 validity and the measure also discriminates between children with and without special 10 education needs (SEN), suggesting good discriminant validity (Van Waelvelde et al., 2012). 11 Whilst both the BHK and SOS seem to be well evaluated for use in Dutch, it is not clear how 12 applicable the measures are to English, which is linguistically and educationally different. 13 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 14 15 educational value and techniques employed in handwriting instruction in different cultures. In 16 the U.K. and U.S., there are considerable variations (within country) of the frequency, 17 duration, and methods of instruction of handwriting (Graham et al., 2008; Stainthorp et al., 18 2006). However, in other countries, such as the Czech Republic, a highly consistent approach 19 is taken, whereby all children are taught the same cursive script (Bartošová et al., 2012; 20 Kučerová & Kucharská, 2018). Such cross-cultural differences in approaches to instruction 21 align with large handwriting style and legibility variability in (e.g., cursivity and specific slant 22 of the script) in countries where inconsistent approaches are taken (e.g., the U.K. and US; 23 Asher, 2006), and greater uniformity in others that adhere to a consistent approach (e.g., the 24 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, 25

1 1984) or test type. For example, measuring the legibility of cursive writing is likely to be less 2 valued in countries allowing flexible (or variable) approaches to handwriting instruction, and 3 where cursive handwriting may not be taught or used by at all, than in countries where this is 4 the norm (Kučerová & Kucharská, 2018). It is important that handwriting legibility 5 assessments have ecological validity for the populations for which they are intended; at the 6 same time, however, an interesting goal should be to develop scales evaluating components 7 of handwriting which generalize for direct comparisons across script types, or at least within 8 script families, as in alphabetic orthographies using Latin letters, for example. However, to 9 date, no such attempt has been made. 10 Unfortunately, since Rosenblum et al.'s (2003) review, - to our knowledge - there 11 has been little development or psychometric evaluation of analytic measures capable of 12 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.

17 The SaHLT also offers a tool enabling the further understanding of the concurrent 18 relationship between spelling and handwriting. Theoretically, van Galen's (1991) 19 psychomotor model makes some predictions about the relationship between spelling and 20 handwriting production with spelling processes being assumed to cascade down, exerting 21 influence on the motor processes associated with handwriting. Experimental work examining 22 the effects of spelling complexity and spelling ability on the dynamics or fluency of 23 handwriting production (Kandel & Perret, 2015; Kandel et al., 2017; Suárez-Coella et al., 24 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 25

1 spelling and handwriting legibility. Caravolas et al. (2020) found spelling ability to be the 2 strongest predictor of handwriting legibility over and above that of schooling experience and 3 reading ability. Pritchard et al. (2021) examined this relationship in younger children and 4 took the view that the quality of motoric representations realised in handwriting legibility 5 holds an important role in later spelling development. This prediction is in the opposite 6 direction of the prediction from van Galen's (1991) model. The authors found a small but 7 significant relationship between handwriting legibility at 6 years and word spelling 6 months 8 later. Gosse et al. (2021) examined the longitudinal relations between spelling, handwriting 9 fluency, and legibility. Using an experimental task in which spelling accuracy, handwriting 10 fluency, and handwriting legibility were derived from single words, the authors found that 11 spelling accuracy was related to the development of fluency, yet no association between 12 spelling accuracy and legibility was found. To date then, the little work on the relationship 13 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 14 15 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 16 17 (SaHLT), we sought to develop and evaluate a concurrent measure of spelling and 18 handwriting which can be used in the classroom, clinic, and in the lab.

19

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.

1 In line with current understanding of spelling development, we selected words graded 2 in difficulty, reflecting the range of spelling patterns that are taught in the British English 3 curriculum over the primary years. The resulting measure of ten sentences (62 words) is 4 demonstrably sensitive for use with children aged 7 to 12 years of age (Class Years 2 to 7) (5 Caravolas et al., 2005). The sentences range in length from 4 to 8 words, and word length 6 varies from 1 to 9 graphemes (i.e., letters or letter groups corresponding to a single sound or 7 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 8 9 orthographic structure. The choice of targeted spelling units making up the word and 10 sentence structures was informed by level-appropriate national curriculum guidelines for Key 11 Stages 1 and 2 (Department for Education, 2014). Thus, the earlier items are shorter and 12 easier from a spelling and grammatical point of view, while the later items include more 13 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 14 15 (<musician's>). In addition, phonological complexity of the words was manipulated in terms 16 of number of phonemes, number of syllables, and consonant clusters throughout the test. 17 Phonologically, the words ranged from the single phoneme/single syllable word, through 18 relatively simple one- or two-syllable words containing only two to four phonemes, to rather 19 complex, multisyllabic words like <performance> - /pəfo:mənts/.

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

1 spacing, and line alignment – which tap the perceptual-motor processes used in children's 2 handwriting production. Letter formation measures presumably rely heavily on letter 3 representations and their associated motor programmes, and they assess the accuracy of each 4 letter's motor programme. Specifically, knowledge of the letter is measured by the letter's 5 form, orientation, and consistency of its angle and size. letter spacing and word spacing 6 dimensions reflect the visuospatial attributes at the inter-letter and inter-word levels by 7 measuring the degree and consistency of the spacing. Finally, line alignment taps both 8 visuospatial and motor control processes. It follows that visuospatial planning is required for 9 the child to position writing on the line, whilst motor control processes contribute to 10 planning, initiating, and maintaining the letters on the line. Therefore, letters that float 11 above/cross below the line may indicate difficulties in visuospatial planning and/or motor 12 control.

13 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 14 15 dimensions of legibility using a Likert-scale from one (very illegible) to five (highly legible). 16 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 17 18 scale with theoretically driven dimension choices to capture specific processes used in 19 handwriting production at the letter and word level. The characteristics of our dimensions 20 were developed by first deconstructing the legibility characteristics of the underlying 21 handwriting processes (e.g., Graham et al., 2006; van Galen, 1991). For example, a letter that 22 is well formed and recognisable suggests a good representation of its specific allographic 23 form (e.g., lower case, cursive), as well as the ability to correctly execute its motor 24 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 25

14

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.

8 In some cases, we combined characteristics of underlying processes – which are 9 assumed to be different - into one dimension. Specifically, in adults, sizing is assumed to be 10 an independent process to allograph selection (van Galen, 1991). However, in children, 11 producing letters that are proportionately larger (< red >) or smaller (< red >) than other letters 12 within a word (sizing errors) could reflect either poor letter formation knowledge, 13 unstable/unconsolidated representations of specific allographs, problems at the motor 14 programming level, or poor motor control. During the initial scoring, we found that children 15 who made sizing errors often did so in a consistent manner for specific letters across words 16 and sentences within their response. Therefore, these errors were more likely the result of 17 poor letter formation knowledge and so letter sizing was considered a feature of the Letter 18 Form dimension.

19 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

8

9 **Participants**

10 A total of 1,461 children completed the SaHLT as part of three projects between 2015 11 and 2020 (please note that some of the monolingual participants in Caravolas et al., 2020) are 12 also included in this sample). This sample size was determined to be sufficient for the 13 analysis techniques proposed below. The children attended school Class Years 3–6 and came 14 from schools located in North Wales (see Table 1). Relatively fewer children in Class Year 6 15 took part as they completed the measures used in this sample as part of a smaller project. All 16 schools taught predominantly through the medium of English and used English for day-to-17 day communication; Welsh was formally taught as a second language. Of this sample, 94% 18 spoke English as their home language, 0.3% spoke Welsh, and the remaining 5.7% spoke a 19 variety of other languages. A brief survey revealed that the group was predominantly right-20 handed (88%), with 11.8% being left-handed and 0.2% being ambidextrous. The proportion 21 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 22 23 very similar to the average of the local area (18.8%) and of Wales (20.1%; Welsh 24 Government, 2018).

1	Participants completed the SaHLT along with measures of spelling, and teachers
2	completed the handwriting proficiency screening questionnaire (HPSQ; see next section). For
3	all participants, head teachers provided full written informed consent, while parents provided
4	opt-out informed consent. This consenting procedure was used as testing was carried out in
5	the classroom and constituted typical classroom activity. This study was conducted in
6	accordance with the British Psychological Society Code of Ethics and Conduct.
7	[Table 1 here]
8	
9	Measures
10	Spelling and Handwriting Legibility Test (SaHLT)
11	We offer a comprehensive description of this measure in the introduction and Online
12	Resource 1 and so we only briefly describe it here. The SaHLT comprises spelling and
13	handwriting legibility which can be administered to children in groups or individually. The
14	spelling subtest includes ten sentences dictated aloud, and children are asked to write the
15	sentence exactly as they hear them on lined paper in their normal handwriting. Each sentence
16	was repeated at regular intervals until the last child had finished writing. Accuracy was
17	measured by awarding one point for every correctly spelled word. The maximum score was
18	62.
19	Handwriting legibility was scored separately, and without regard for spelling
20	accuracy, by a trained research assistant or teacher. The handwriting legibility scoring criteria
21	were applied at the sentence level to the written productions. Each of the ten sentences was
22	scored according to the criteria of each dimension (letter formation, letter spacing, word
23	spacing, and line alignment) on a five-point Likert scale ranging from 1 to 5 (very illegible to
24	highly legible). The scale was applied aggregating over the relevant units (letters, words,
25	respectively) for each of the dimensions in each sentence. An overall legibility score for each

26 sentence was obtained by summing across the four-dimension scores (possible maximum

18

1 being 20 points). These scores were then averaged over the ten sentences leading to a mean

2 maximum possible score of 20 for the test overall. A higher score indicates greater legibility.

3 Word Spelling

4 Word spelling ability was measured using the word spelling subtest from the Wide 5 Range Achievement Test IV (WRAT-IV; Wilkinson, & Robertson, 2006). This measure was 6 adapted for group administration, following the published guidelines closely. Children 7 spelled to dictation 13 alphabet letters followed by words graded in difficulty. Children in 8 Year 6 completed all 42 words. Children in all other class year groups completed 36 words. 9 The cut-off for younger children, corresponds to a standard score of 145 for Class Year 5 10 children in the U.K. It was expected that most children were unlikely to exceed this score. 11 Each correctly spelled word was awarded one point and scoring was discontinued after 10 12 consecutive errors. The internal reliability of this measure was good ($\alpha = .89$).

13 Handwriting Proficiency Screening Questionnaire (HPSQ)

14 Teacher assessments of handwriting were sought using the Handwriting Proficiency 15 Screening Questionnaire (HPSQ; Rosenblum, 2008). The questionnaire includes 10 questions 16 which measure wellbeing, handwriting fluency, and handwriting legibility. Teachers were 17 asked to observe children writing and to answer 10 questions, such as 'Is the child's 18 handwriting unreadable', using a five-point Likert scale from 0 (*Never*) to 4 (*Always*). A 19 higher score indicates lower handwriting ability. The internal reliability of this measure was 20 good ($\alpha = .89$).

21 **Procedure**

22 Children completed the measures in specially prepared booklets in typical classroom 23 conditions. The measures were delivered over two sessions not exceeding 60 minutes. All 24 sessions were conducted by a team of at least three trained research assistants who 25 maintained good oversight of children's work and of their compliance with the instructions.

1	All children completed all measures. Teacher's ratings on the HPSQ were taken from 107
2	children. Those scored were selected based on (a) teacher availability to complete the
3	observation and scoring (b) child availability to be observed and scored.
4	Results
5	To meet the first aim of this study, to evaluate the reliability and validity of each of the
6	two scales of the SaHLT, each in its own right, we undertook a rigorous psychometric
7	evaluation of the spelling and handwriting legibility subtests. Second, to meet our overarching
8	aim to investigate the concurrent relationship between the spelling and handwriting scales of
9	this assessment tool, we undertook multi-group factor analysis
10	Evaluation of the Psychometric Properties of the Spelling and Handwriting Legibility
11	Subscales
12	Developmental Sensitivity
13	We sought to test the developmental sensitivity of the spelling and handwriting
14	subtests of the test. We did so by comparing the class year (i.e., school grade) means on the
15	total spelling and handwriting scores, respectively (see Table 2 for means and standard
16	deviations for each class year group). As expected, there were moderate-to-large differences
17	in spelling accuracy between class years, $F(3, 1455) = 70.83$, $p < .001$, $\eta_p^2 = .13$. Post-hoc
18	comparisons with Bonferroni corrections revealed spelling accuracy significantly increased
19	with each grade. That is, children in Class Year 6 were more accurate than those in Class
20	Year 5 ($p < .001$), who were more accurate than those in Class Year 4 ($p < .001$), who, in
21	turn, were more accurate than those in Class Year 3 ($p < .001$). There were also moderate-to-
22	large differences in handwriting legibility between class years, $F(3, 1456) = 93.41, p < .001,$
23	$\eta_p^2 = .16$. Post-hoc comparisons with Bonferroni corrections showed children in Class Year 6
24	had significantly better handwriting legibility than children in Class Year 5 ($p < .001$), who

1	had significantly better handwriting legibility than children in Class Year 4 ($p = .021$), who,
2	in turn had significantly more legible writing that children in Class Year 3 ($p < .001$).
3	[Table 2 here]
4	Internal Reliability
5	We assessed internal reliability by examining Cronbach's alpha as a function of each
6	year group and in total for the sentence level scores on both the spelling and total handwriting
7	legibility subtests. The estimates reported in Table 3 show excellent reliabilities for both
8	spelling and handwriting legibility subtests.
9	[Table 3 here]
10	Test-retest Reliability
11	A representative sub-sample of the total number of children in Class Years 3, 4, and 5
12	completed the SaHLT a second time, approximately seven months after the initial
13	assessment. Responses at the two time points were scored by the same assistant. Test-retest
14	reliability was assessed by calculating the intra-class correlation (ICC) between children's
15	performance on the sentences administered at times one and two. Table 4 reports the ICC and
16	associated 95% confidence intervals for each Year Group on the spelling accuracy (left
17	panels) measures of the SaHLT. As can be seen, the spelling accuracy measure had good to
18	excellent reliability (Koo & Li, 2016). The ICCs of the handwriting subscale demonstrate this
19	measure had moderate-to-good test-retest reliability (Koo & Li, 2016).
20	[Table 4 here]
21	Inter-rater Reliability
22	For spelling, a research assistant and the first author scored a representative sub-
23	sample of the total children tested in Class Years 3 – 6. Inter-rater reliability was evaluated
24	using a two-way random effects ICC. The ICCs and 95% confidence intervals for each year
25	group are reported in Table 5. The ICCs report a very high level of agreement between raters
26	(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).

6

7

Convergent Validity

[Table 5 here]

8 We analysed the convergent validity of the Spelling subtest by comparing it with a 9 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 10 11 Spelling was statistically significant (p < .001) and strong across all class years (Class Year 3) 12 r = .76; Class Year 4 r = .86; Class Year 5 r = .82; Class Year 6 r = .77), with an overall 13 correlation of r = .84. This demonstrates good convergent validity for spelling. 14 We measured the convergent validity of the SaHLT's handwriting legibility measure 15 by examining its association with teacher ratings on the Handwriting Proficiency Screening 16 Questionnaire (HPSQ; Rosenblum, 2008) amongst 107 children (33 children in Class Year 3, 17 41 children in Class Year 4, and 33 children in Class Year 5) from the sample. Unfortunately, 18 due to logistical issues, we were unable to ascertain teacher responses on the HPSO for 19 children in Class Year 6. The HPSQ is a valid and reliable screening questionnaire for 20 teachers designed to measure teacher perceptions of a child's handwriting legibility, fluency, 21 and well-being. Well-being in this context is described as including children's willingness to 22 write and capturing fatigues whilst writing (Rosenblum, 2008). In this questionnaire, higher 23 ratings from teachers indicate poorer handwriting. The correlations between the SaHLT 24 Handwriting scores and the HPSQ were statistically significant (p < .001) and large in size.

- 25 The correlations ranged from r = -.61 in Class Year 3, r = -.72 in Class Year 4, and r = -.64 in
- 26 Class Year 5, giving a total correlation of r = -.66 across the sample. The size of these

1 correlations is smaller than those reported for the spelling test, but these correlations were 2 expected as the HPSQ measures aspects of handwriting (e.g., fluency and wellbeing) along 3 with legibility whereas both the spelling measures simply measure spelling accuracy. In sum, 4 these correlations demonstrate good convergent validity of the handwriting subscale. 5 **Factor Structure of Concurrent Spelling and Handwriting Measurement** 6 We sought to examine the factor structure of a concurrent measure of spelling 7 accuracy and handwriting legibility. To do so, we initially ran bivariate correlations between 8 all sub-measures of the SaHLT. We split the spelling measure into three sub-measures: 9 accuracy of function words (n = 22 words; $\alpha = .85$), accuracy of morphologically simple 10 (uninflected; n = 20 words; $\alpha = .93$), and accuracy of morphologically complex (inflected; n = 20 words; α = .91) words. For handwriting, this included letter formation (α = .95), letter 11 spacing ($\alpha = .94$), word spacing ($\alpha = .94$), and line alignment ($\alpha = .95$). The correlations 12 13 reported in Table 6 show strong correlations between the spelling subtest measures. The 14 correlations between the handwriting subtest measures were also strong, but slightly weaker 15 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. 16 17 [Table 6 here] 18 We used confirmatory factor analysis (CFA) to examine the factorial structure of the 19 SaHLT (factorial validity) and test whether this structure held over the year groups. The 20 purpose of doing so was to test whether it was possible to measure the two constructs of 21 spelling accuracy and handwriting legibility concurrently in children aged 7-11 years old and 22 then to examine the relationship between the two constructs. In our initial CFA, the model

23 was applied to the whole sample (N = 1,460). The model was run using direct ML estimation

24 in Mplus 8.4 (Muthén & Muthén, 2019). Direct ML was favoured given a small amount

25 (0.78%) of missing data completely at random (MCAR). MCAR status was further

1 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 $\chi^2(4) = 5.13$, 2 p = .274. Direct ML is preferred as it uses all the available data (Brown, 2015). Data were 3 4 approximately normally distributed and there were no significant outliers. 5 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 6 7 loaded onto one spelling factor. All the handwriting legibility measures (letter formation, 8 letter spacing, word spacing, and line alignment) were loaded onto a second, handwriting, 9 factor. Furthermore, the residual variances between letter formation and letter spacing were 10 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 = 11 12 .98 and TLI = .96, and all loadings and correlations were significant. Overall, there were high 13 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. 14 15 The single-group model was then applied to a multigroup CFA (MGCFA), which 16 allows the testing of measurement invariance to assess whether the measures were 17 comparable across the groups. That is, it ensures the model and the indicators were valid 18 measures across all class year groups (Brown, 2015; Milfont & Fischer, 2010). To assess 19 measurement invariance a stepwise procedure of successively more restricted models was 20 used as recommended by Brown (2015), Milfont and Fischer (2010), and Vandenberg and 21 Lance (2000). Using this procedure, the more constrained model is a nested version of the 22 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 23 24 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).

3 Table 7 shows that constraining the factor structure yielded configural invariance 4 equal across groups and produced an acceptable fit. This model acted as the baseline for the 5 further, more constrained models. In the next analysis, the equality of factor loadings (metric 6 invariance) was tested between class years by constraining factor loadings to be equal across 7 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 8 9 there any change in the CFI value. In the next model, the intercepts of the indicators were 10 constrained to be equal across all class years (scalar invariance). This model did not differ 11 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. 12

13 Invariance across indicator residuals (differences between measurement error between 14 groups) was not tested as it was deemed to be overly restrictive given there were no 15 theoretical or methodological reasons to expect errors to be equal across class years (see 16 Brown, 2015). Similarly, structural invariance was not tested as performance on indicators 17 did change developmentally (see earlier section on developmental sensitivity). The analyses 18 demonstrate the current solution has measurement invariance, indicating the spelling and 19 handwriting measures were separable and valid measures of their respective constructs across 20 the four class years. The path diagram of the final model the accompanying unstandardized 21 and standardised factor loadings and indicator residual variances for each class year are 22 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.

4





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

- 11
- 12 13

[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

1 whether it was possible to robustly measure spelling and handwriting legibility concurrently. 2 We found that the spelling test was a suitable measure of English spelling, and it was possible 3 to robustly measure spelling and handwriting legibility concurrently in mid-primary-school-4 aged children. Specifically, the concurrent measure demonstrated sensitivity to age related 5 increases in both spelling accuracy and handwriting legibility. Furthermore, both spelling and 6 handwriting measures had good internal and test-retest reliability, and handwriting had good 7 inter-rater reliability. We also found that the measures of both spelling and handwriting 8 correlated well with respective established tests of spelling and handwriting, confirming 9 convergent validity. In addition, we sought to test the structure and inter-relationships 10 between the constructs. We found the test measured performance on two factors: spelling and 11 handwriting legibility constructs. These were independent of one another but were 12 interrelated. This model applied equally across the four class years, suggesting good stability. 13 To our knowledge, this is the first assessment of a concurrent measure of sentence 14 spelling and handwriting legibility. We argue that it is theoretically informative, cost and 15 time effective to concurrently measure spelling and handwriting. The SaHLT was able to 16 capture with some sensitivity age related changes in spelling and handwriting legibility. 17 Scores on the sentence spelling test increased as a function of grade, suggesting that the 18 graded increases in the linguistic complexity of the words presented in the sentences 19 adequately captured age-related increases in spelling ability. This supports the view that a 20 sentence spelling task can offer insights about children's single word spelling ability that are 21 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

1 that they capture variations in instructional approaches to handwriting. It follows that in areas 2 where there is a highly consistent instructional approach to handwriting (e.g., Czech 3 Republic; Bartošová et al., 2012) typically developing children are likely to reach their 4 optimum legibility at a younger age. This, in turn, reduces the ability to capture variance in 5 older children. Whereas, in areas where there is less consistency in the instructional 6 approaches to handwriting (e.g., U.S.; Graham et al., 2008), children may not reach their 7 optimum legibility as early and so it easier to capture variance for longer periods of time. 8 Whilst direct cross-education-system studies – which would test this theory – are lacking, 9 indirect evidence is consistent with the above hypothesis. For example, early plateaus of 10 handwriting legibility are reported in systems with a highly structured approach to 11 handwriting (e.g., Belgium; c.f., Gosse et al., 2021) but not in systems with an unstructured 12 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 13 previous methods of estimating handwriting legibility have not had adequate reliability (c.f., 14 15 Rosenblum et al., 2003). We argue that generalizable handwriting legibility measures should be developed and validated across writing systems to test these theories. 16

17 Rosenblum et al. (2003) highlighted several issues regarding the proper assessment of 18 handwriting; these included (a) a lack of clarity in scope of the measures, (b) how practical 19 the measure is to administer, and (c) a lack of robust evaluation of handwriting measures. 20 Unfortunately, since Rosenblum et al. (2003) little progress on dealing with these issues 21 seems to have been made in relation to handwriting legibility. The current study contributes 22 new insights to these issues. The SaHLT can be used for both screening via whole class 23 administration and for individual assessment. We have demonstrated that the test can be used 24 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.

9 In what may well be the most comprehensive psychometric evaluation of a 10 handwriting legibility measure to date, we found measuring handwriting legibility across four 11 dimensions to be a reliable and valid test of handwriting legibility amongst children aged 7 – 12 11 years old. The testing and reporting of the reliability and validity of handwriting legibility 13 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 14 15 measure fared well. On inter-rater reliability, the SaHLT handwriting legibility subtest 16 performed similarly to the BHK (Hamstra-Bletz et al, 1987) and better than the SOS 17 (Waevelde et al., 2005) and Ziviani and Elkins (1984). Similarly, on test-retest reliability, the 18 current measure was more reliable than the SOS (Van Waevelde et al., 2005). In addition to 19 reliability, we evaluated several types of validity and found the measure to be valid. Our 20 evaluation of the handwriting subtest's convergent validity found it to be weaker than the 21 BHK's (Hamstra-Bletz et al, 1987). However, it is important to note that we validated our 22 subtest with a measure that tapped both fluency and legibility (the DASH; Barnett et al., 23 2007) whereas Hamstra-Bletz et al. (1987) correlated the BHK scores with direct teacher 24 ratings of legibility. As such, this discrepancy is likely to stem from the choice of the validation measure used. 25

1 We were also able to assess the construct validity of the spelling and handwriting sub-2 measures using factor analyses. Crucially, these techniques also afforded us the opportunity 3 to directly test the question as to whether it is possible to concurrently measure spelling and 4 handwriting legibility as separable constructs. The factor analysis confirmed the presence of 5 two separable constructs that, moreover, were developmentally stable for 7–11 year olds. 6 Furthermore, the correlations between the constructs revealed that spelling and handwriting 7 legibility are moderately related, and that this relationship is of a similar strength for children 8 aged 7–11 years. However, we did observe a noticeably smaller correlation in the older class 9 year group of children (10–11-year-olds) in this sample. This correlation may have weakened 10 due to the decoupling of spelling and handwriting legibility as the skills become increasingly 11 automatised in older children (Gosse et al., 2021; McCutcheon, 1996). An alternative - but 12 not as compelling – explanation is that the weaker correlation may reflect the smaller n (n =13 82) in this class year.

14 It may be argued that the observed relationships between spelling and handwriting 15 legibility measures stem from shared method variance, as both measures are derived from the 16 same production. However, in previous research with the SaHLT, we found that a significant 17 relationship obtains between other measures of spelling (e.g., single word spelling) and the 18 SaHLT handwriting legibility dimensions used in the current study (Caravolas et al., 2020). As 19 such, it is unlikely that shared method variance accounts for the relationships found here. The 20 presence of a significant concurrent relationship between spelling and handwriting legibility 21 adds weight to the existing literature also reporting a relationship between these skills in typical 22 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

1 influence this relationship, including the curricular emphasis that is placed on handwriting 2 skills in different educational, cultural, and linguistic systems. Such direct comparisons 3 would allow us to further understand the language-, orthography- and culture-general 4 relationships between spelling and handwriting, and the between-study (and country) 5 differences in the timing of plateauing of handwriting legibility, which seems to occur 6 earlierin some areas (e.g., Belgium) than in others (e.g., U.K.). The current test could be 7 readily adapted to other languages; for example, a parallel spelling measure has already been 8 created in Czech (Caravolas et al., 2005). Furthermore, the current handwriting legibility 9 dimensions - created for assessing legibility in a highly variable educational context - could 10 well be a suitable measure for handwriting legibility components which are universal across 11 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).

17 In sum, we sought to investigate the concurrent relationship between handwriting and 18 spelling. We found that a sentence spelling task with stimuli varying in their linguistic 19 complexity is sensitive to grade/age-related differences, and offers a reliable and valid 20 alternative to single word spelling, the typical assessment method in English. Furthermore, 21 our study demonstrated that it is possible to robustly measure spelling and handwriting 22 legibility concurrently, and, that they are separable, but related abilities. The measure we 23 evaluated in this paper offers a cost and time saving measure of both constructs 24 simultaneously. Furthermore, this soon-to-be freely available measure (for further

30

- 1 information please see: <u>http://www.eldel-bll.uk/sahlt/</u>) can be used in future investigations of
- 2 the relationship between spelling and handwriting legibility.

3

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. <u>https://doi.org/10.1037/a0019318</u>.
- 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. <u>https://doi.org/10.1016/S0010-9452(08)70499-4</u>.
- 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. <u>https://doi.org/10.1016/j.humov.2015.08.011</u>
- 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., Maněnová, 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. <u>https://doi.org/10.1177/1534508410380136</u>
- 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. <u>https://doi.org/10.1027/1016-9040.9.1.3</u>

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. <u>https://doi.org/10.1207/S15328007SEM0902</u>

- 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. <u>https://www.gov.uk/government/publications/national-</u> <u>curriculum-in-england-framework-for-key-stages-1-to-4/the-national-curriculum-in-</u> <u>england-framework-for-key-stages-1-to-4</u>
- 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-difficultiesamongst-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. <u>https://doi.org/10.1080/00220679809597574</u>.
- 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. <u>https://doi.org/10.1007/s11145-007-9064-z</u>
- 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. <u>https://doi.org/10.1207/s15326942dn2901_4</u>
- 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. <u>http://dx.doi.org/10.1016/j.jcm.2016.02.012</u>
- 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. <u>https://doi.org/10.1007/s11145-018-9929-3</u>
- Masterson, J. J., & Apel, K. (2010). The spelling sensitivity score: Noting developmental changes in spelling knowledge. Assessment for Effective Intervention, 36(1), 35–45. <u>https://doi.org/10.1177/1534508410380039</u>
- McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review*, 8(3), 299–325. <u>https://doi.org/10.1007/BF01464076</u>
- 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. <u>https://doi.org/10.21500/20112084.857</u>
- 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. <u>https://doi.org/10.3389/fpsyg.2013.00818</u>
- 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. <u>https://doi.org/10.1080/10888438.2020.1778705</u>.
- 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. <u>https://doi.org/10.1007/s11145-017-9762-0</u>
- Rosenblum, S. (2008). Development, reliability, and validity of the Handwriting Proficiency
 Screening Questionnaire (HPSQ). *The American Journal of Occupational Therapy*,
 62(3), 298–307. <u>https://doi.org/10.5014/ajot.62.3.298</u>
- 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. <u>https://doi.org/10.1016/S0167-9457(01)00033-1</u>
- 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. <u>https://doi.org/10.3389/fpsyg.2020.00319</u>.
- 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. <u>https://doi.org/10.1016/j.actpsy.2017.12.001</u>
- 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. <u>https://doi.org/10.1080/02643294.2017.1337630</u>
- 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. <u>https://doi.org/10.1080/0013191840360304</u>

Description	of the	Evaluation	Sample
-------------	--------	------------	--------

		n by g	gender	Age (months)			
Class year	N	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		

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

		п	М	SD	Range	Max score
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	
		Handwrit	ing legibil	ity		
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.

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

	S	pelling	g accura	асу	Handwriting legibility					
-	Delay ^a	п	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]		

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

Note. ^aAverage number of months between time 1 and time 2. ICC = Two-way mixed effects intra-class correlation. CI = Confidence interval.

		Spelling		Han	dwriting
	п	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]

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

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

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.

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.

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

	Unstand	dardised				Standa	rdised			
			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.