Article

LANGUAGE TESTING

Language Testing
2020, Vol. 37(3) 435–452

DOI: 10.1177/0265532220909310
journals.sagepub.com/home/ltj
SAGE

An innovative measure of orthographic processing: Development and initial validation

Yi-Jui Iva Chen
University of South Florida, USA
University of California, USA

Mark Wilson

Robin C. Irey
University of California, USA

Mary K. Requa
University of California, USA
San Francisco State University, USA

Abstract

Orthographic processing – the ability to perceive, access, differentiate, and manipulate orthographic knowledge – is essential when learning to recognize words. Despite its critical importance in literacy acquisition, the field lacks a tool to assess this essential cognitive ability. The goal of this study was to design a computer-based assessment of orthographic processing and investigate its psychometric properties. The rationale for designing specific items was discussed, methods used to separate orthographic processing from word recognition and spelling ability were presented, and item suitability was examined. Person separation reliability was .91 for this assessment. Validity evidence was gathered and reported.

Keywords

Assessment development, literacy skill, orthographic processing, test validation

Corresponding author:
Yi-Jui Iva Chen, Rightpath Research and Innovation Center, University of South Florida, College of Behavioral and Community Sciences, 13301 Bruce B. Downs Blvd. MHC 1704, Tampa, FL 33612, USA.
Emails: chen16@usf.edu; iva811@berkeley.edu

Automatic word recognition allows the reader to allocate cognitive resources to higher-level reading processes, such as attending to the syntactic and semantic structure of language and evaluating and integrating text (Perfetti, 2007; Perfetti, Landi, & Oakhill, 2005). Given the role of automatic word recognition in reading proficiency, it is essential to understand the various processes that underlie and influence the development of this important skill. Although phonological processing is the primary determinant of word identification (Adams, 1990; Share, 1995, 1999), it is becoming increasingly clear that phonological skills alone cannot explain the development of skilled word recognition (Cunningham, Perry, & Stanovich, 2001; Cunningham, Nathan, & Raher, 2011; Share, 1995; Stanovich & West, 1989). Studies have shown that orthographic skills contribute unique variance in word recognition after controlling for phonological skills (Stanovich & West, 1989; Olson, Wise, Conners, Rack, & Fulker, 1989), age, and non-verbal intelligence (Cunningham & Stanovich, 1990). It is likely that orthographic skills represent an additional, independent, and necessary component of word recognition (Apel, 2009; Apel, Thomas-Tate, Wilson-Fowler, & Brimo, 2011; Berninger, Cartwright, Yates, Swanson, & Abbott, 1994; Berninger, Yates, & Lester, 1991).

As proposed by current major cognitive theories of word recognition (Coltheart, 2005; Ehri, 2005, 2014; Harm & Seidenberg, 1999; Perfetti, 2007; Seidenberg & McClelland, 1989; Share, 1995, 2004), word recognition requires well-specified orthographic representations that are linked securely to semantic and phonological information (McClung, O'Donnell, & Cunningham, 2012; Perfetti, 2007). In addition, these theories suggest that a limited ability to process orthographic representations of words leads to reading difficulty, whereas a strength in this ability distinguishes skilled from unskilled readers. This ability is referred to as *orthographic processing*: the processing ability to perceive, access, differentiate, correct, and arrange orthographic representations and orthographic patterns.[1]

Existing assessments of orthographic processing

Despite the importance of orthographic processing, the field lacks a comprehensive measure of this skill. This might be owing to the challenge of constructing "pure" and "accurate" measures of orthographic processing (Wagner & Barker, 1994). The most widely used assessments of orthographic processing are orthographic choice tasks, homophone choice tasks, wordlikeness choice tasks, and rapid letter naming tasks (see Apel, 2011 for a comprehensive review).

Olson et al (1985) and Stanvoich and West (1989) used orthographic choice tasks to measure orthographic processing. In orthographic choice tasks, respondents choose the correct spelling between two options that sound alike (e.g., *dream-dreem* or *rain-rane*). Slightly different from orthographic choice tasks, homophone choice tasks (Stanovich & West, 1989) involve a question that is read orally to respondents (e.g., which is a fruit?) and then respondents choose an answer between two homophones (e.g., *pair-pear*). Because the options sound alike, phonological decoding is not the cause of performance difference on these two types of tasks. Rather, orthographic skills explain performance difference on the orthographic choice and homophone choice tasks.

In a wordlikeness choice task, respondents are required to choose a pseudoword with a permissible letter combination (e.g., *filv-filk*). This task has been widely used in literacy

research (e.g., Treiman, 1993; Stanovich & Siegel, 1994; Cassar & Treiman, 1997; Tsung, Zhang, Hau, & Leong, 2017) as a measure of knowledge of orthographic patterns. In a rapid letter naming task, respondents read letters aloud and the time spent naming the letters is recorded. For example, Liu et al (2017) used the Letter Naming Facility task from the Kaufman Test of Education Achievement (Kaufman & Kaufman, 2014) to measure orthographic processing.

Nevertheless, these assessments are limited in several ways. Orthographic choice and homophone choice tasks actually measure word recognition ability or spelling ability, rather than orthographic processing. Because respondents are required to choose the correct spelling of a word between two options that sound alike, responses heavily depend on knowledge of English spelling, rather than on the ability to perceive, access, differentiate, correct, and arrange orthographic representations and orthographic patterns. Despite this confounding issue, the orthographic choice task and the homophone choice task are still widely used in literacy research (e.g., Cunningham, 2006; Deacon, Commissaire, Chen, & Pasquarella, 2012). Rapid letter naming tasks are also limited because they focus on single letters, and thus fail to measure an individual's ability to manipulate orthographic patterns. The wordlikeness choice tasks use pseudowords to address the confounding issue between the processing of orthographic patterns and word recognition and spelling ability; as such, this type of task is the best attempt thus far to measure orthographic processing. However, it is limited because it focuses on processing spelling patterns alone and ignores fundamental orthographic processing skills (i.e., the ability to perceive or access orthographic representations). Moreover, evidence has shown that the worklikeness choice tasks might also involve phonological skills (Hagiliassis, Pratt, & Johnston, 2006). Measuring orthographic processing via the orthographic choice and homophone choice tasks or solely with a wordlikeness choice task or rapid naming task are not ideal strategies (Burt, 2006; Castles & Nation, 2008; Chalmers & Burt, 2008; Hagiliassis, Pratt, & Johnston, 2006).

An innovative approach to measuring this complex skill should go beyond single letters, cover multiple processing skills, and disaggregate orthographic processing from word recognition and spelling ability. The goal of this study was to develop a computer-based assessment that measures an individual's ability to perceive, access, differentiate, and manipulate orthographic representations and orthographic patterns. The design of the assessment was based on three principles: (a) it must measure the ability to process single letters, letter combinations, and spelling patterns; (b) it must include multiple processing skills; and (c) it must allow for the separation of orthographic processing from word recognition or spelling ability. Item response theory was used to evaluate and calibrate the test items.

Method

Participants

One hundred forty participants (62 boys and 78 girls) were administered the computer-based orthographic processing assessment. Thirty-four of the participants were third-to-sixth-grade English learners in Taiwan. The mean age in years at the time of data collection for each grade in this group was 9.07 ($SD = 0.21$), 10.26 ($SD = 0.24$), 11.26

(*SD* = 0.40), and 12.35 (*SD* = 0.26), respectively. One hundred six of the students were first-to-fifth-grade students from four schools in two suburban cities and one urban city located in northern California. The mean age in years at the time of data collection for each grade in this group was 6.95 (*SD* = 0.03), 8.02 (*SD* = 0.33), 8.76 (*SD* = 0.31), 10.50 (*SD* = 0.19), and 10.64 (*SD* = 0.69), respectively. We recruited the Taiwanese sample because we were interested in conducting a differential item functioning (DIF) study between Taiwanese and American students. This initial DIF investigation provided rich information for the field of orthographic processing and informed the calibration process for the items developed for this assessment, which will be used to inform a future larger-scale validation study.

The Taiwanese participants in this sample received 80 minutes of English instruction every week during the school year beginning in Grade 1 and were not exposed to English outside of school. Students in Grade 3 typically begin to master the basic rules of letter–sound correspondence. Therefore, we recruited older Taiwanese participants to avoid a floor effect. We recruited the English as foreign language (EFL) students in Taiwan and the native English speakers in the USA because we wanted to compare these two unique groups to see if performance on these tasks would be different.

Design of a novel orthographic processing assessment

We based the design of our assessment on Stanovich and West's (1989) conceptualization of orthographic processing. We used Cassar and Treiman's (1997) word list and Cummings' (1988) American English spelling system to design items. The design of our tasks was rooted in information processing theories (Carroll, 1999; Mayer, 1996). Based on these resources, we created perception, access, differentiation, correction, and arrangement tasks that measure an individual's cognitive processing ability to perceive, access, differentiate, correct (addition and removal), and arrange orthographic representations and orthographic patterns.

The items were designed to correspond to these processing domains. Each task included 1–4 practice items and 9–12 actual items. The first four tasks, that is, perception, access, differentiation, and correction (addition, and removal), required a fixed response, whereas the fifth task (arrangement) required an open-ended response. It was hypothesized that the open-ended response task (i.e., the arrangement task) would be more challenging than a fixed response task, because respondents were required to produce a response rather than choose an answer (i.e., the correction task).

The computer software *E-Prime 2.0 Professional* (Psychology Software Tools, 2012) was used to deliver the tasks. *E-Prime* provides millisecond timing to ensure the precision of the data collected and the manipulation of stimuli so that participants' responses reflect the efficiency of their orthographic processing. The instructional statements and questions appeared in New Courier size 24 black font and stimuli and answer choices were presented in Arial size 36 blue font. Appendix A provides detailed information for each item.

Assessing processing. In order to distinguish orthographic processing from word recognition and spelling ability, the items were designed to ensure that knowing the spelling of

a word would not guarantee the selection of the correct answer, and conversely, not knowing the spelling of a word would not guarantee an incorrect answer selection. The first goal was met by not providing the correct spelling to respondents but instead presenting stimuli to activate multiple orthographic representations (e.g., create, recreate, accrete) in respondents' minds. This was achieved by the final two tasks, correction and arrangement. For example, for the provided stimulus "a, e, e, c, r, t" respondents were asked to unscramble the letters provided to make a real word, *create*. The scrambled letter presentation was designed to activate several other words (i.e., create, recreate, accrete) that contain the same letters as the target. Thus, respondents were required to process the information provided within an item, evaluate multiple orthographic representations, and provide the correct response. The second goal was achieved by using pseudowords as stimuli in the first three tasks: perception, access, and differentiation.

Item design

In the *perception* task, respondents were required to perceive orthographic representations by recognizing the presence of letters and distinguishing them from other symbols (e.g., *!2@&*) within a short timeframe. They were asked to identify the number of letters among a group of symbols (e.g., *QK&G*) within five seconds.

In the *access* task, respondents were required to store orthographic representations and then retrieve those stored orthographic representations from memory. To assess this, respondents were asked to identify a letter combination among four options after seeing the target letter combination for three seconds followed by the question "*Which group of letters did you see*?" For example, respondents saw the target, *byt*, for three seconds, the question "*Which group of letters did you see*?" for four seconds, and then four options were presented on the screen and respondents were required to choose the target from the options (*byu, byt, dyt, byk*).

In the *differentiation* task, respondents were required to differentiate permissible and impermissible letter combinations in the American English spelling system. They were required to identify the permissible combination among three other impermissible combinations (e.g., *tought, tiught, cuught*, and *cyught*). Some stimuli were adopted from the word list provided in Cassar and Treiman's (1997) study of orthographic knowledge.

In the *correction* task, respondents were required to add or remove a letter, depending on the item, to create a correctly spelled word, adhering to the American English spelling system (e.g., *lauyout, wamter*, or *bbelly*). As previously mentioned, to disaggregate orthographic processing from spelling and word recognition ability, items for this domain were designed so that knowing the spelling of a word did not necessarily guarantee a correct response. For example, an item that required a letter be removed [*wamter*], was designed so several orthographic representations could simultaneously be activated, such as *warmer, winter, water (correct response)*, and *warder.* Respondents then needed to process these orthographic candidates in their mind to determine which one had only one letter difference.

In the *arrangement* task, respondents were required to rearrange provided orthographic representations, demonstrating that they can rearrange letters, adhering to the American

English spelling system, to make a correctly spelled word. For example, respondents rearranged the stimulus letters *l, l, e, e*, and *v* to form the correct response: *level*.

Given the foundational nature of the first two tasks, we included a time limitation for these tasks to test participants' efficiency and automaticity perceiving letters and storing letters. We set an 80-keystroke limit per item because only five to eight keystrokes were needed per item and we wanted to prevent students from repeatedly pressing random keys. We used uppercase and lowercase letters for the first two tasks because these two tasks were focused on the processing of letters, and orthographic processing skills require automatic processing of both uppercase and lowercase letters. In contrast, the final four tasks focused more on processing spelling patterns rather than individual letters, and because words are typically lowercase in text, we only used lowercase letters for the final four tasks.

Data collection

For the U.S. sample, the orthographic processing data were collected in classroom groups at each participating school site. Four trained doctoral level researchers brought laptops to the classrooms. Each researcher attended to five or fewer participants to ensure the computer-based assessment was conducted appropriately. For the Taiwan sample, the orthographic processing data were collected individually. The Peabody Picture Vocabulary Test-Fourth Edition (Dunn & Dunn, 2012) and the Letter and Word Identification subtest of WJ-III (Woodcock, McGrew, & Mather, 2007), used to test concurrent validity, were administered individually by the four trained doctoral-level researchers.

Validity investigation

We used the methods recommended by Kane (2006, 2010, 2013), the Standards for Educational Psychological Testing (American Education Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), and Wilson (2005) for gathering validity evidence. These methods include (a) reviewing internal structure, (b) examining concurrent validity, (c) examining predictive validity, and (d) considering consequential validity. To examine concurrent validity and predictive validity, two standardized assessments and one within-subject experiment (see below) were included in this study to examine the relation between the orthographic processing assessment and language and literacy skills.

Standardized assessment. The Peabody Picture Vocabulary Test-Fourth Edition (PPVT: Dunn & Dunn, 2012) and the Letter and Word Identification subtest of WJ-III (Woodcock et al., 2007) were used to validate this orthographic processing assessment. The PPVT-4 was designed to assess both children's receptive and expressive vocabulary and includes 228 items. In this study, the PPVT-4 was used to measure receptive oral vocabulary. The test developers report that for respondents aged 7–10 years, the test-retest reliability was .91, the split-half reliability ranged from .90 to .95, and Cronbach's alpha ranged from .94 to .97. The Letter and Word Identification subtest of the WJ-III was designed to assess children's ability to recognize letters and words in print. It has 75 items, including letters (e.g., *P, E*), single-syllable words (e.g., *must*), and polysyllabic

words (e.g., *domesticated*). The reliability of this test for respondents aged 7–10 years ranged from .93 to .97 (McCrew, Schrank, & Woodcock, 2007).

It was expected that the orthographic processing assessment would positively correlate with each of these assessments because all cognitive related skills are positively correlated owing to the common factor of general intelligence. However, the correlation between the orthographic processing assessment and receptive oral vocabulary should be lower than the correlation between the orthographic processing assessment and word recognition ability because receptive oral vocabulary is not involved in orthographic representations.

Within-subject experiment. A within-subject computer-based experiment was conducted to test the validity of the orthographic processing assessment. In this experiment, we used a spelling production task to document participants' ability to learn new spellings both immediately after exposure and two days after exposure, providing two data points representing spelling acquisition. It was designed to examine whether or not participants were able to deduce spelling patterns and make orthographic analogies to learn new spellings; for example, whether participants were able to deduce and use the rime similarity between a known word (*rain)* to learn the spelling of an unknown word (*vain*).

More specifically, the computer-based experiment consisted of three stages. Prior to Stage 1, participants were told that they were going to learn new words and that they would be quizzed on the spelling of the words after the session. They were also told that they would see each new word three times for five seconds. During Stage 1, participants selected, from four options, the picture which best corresponded with the meaning of each base word. Participants were required to press a number key (1–4) to answer. This procedure was designed to verify that participants knew the base words and to activate them in their mind. Because the base words were words that the participants already knew, they could use them to make orthographic analogies to learn unfamiliar words. During Stage 2, participants were shown the unknown words and pictures corresponding to their meaning as well as the base words and corresponding pictures. Finally, in Stage 3, participants completed a spelling posttest both immediately after the experiment and two days after the experiment. During the spelling task, participants saw pictures of the previously unknown words and used a pencil to spell the corresponding words. The detail of the within-subject experiment can be found in Chen, Cunningham, Rabe-Hesketh, Hinshaw, and Irey (2019).

It was hypothesized that participants with advanced orthographic processing skills would be better at using orthographic analogies to acquire spelling during the experiment. To test this, we conducted regression analysis to examine whether our researcher-designed orthographic processing assessment was a significant predictor of participants' spelling acquisition after controlling for their oral vocabulary and word recognition ability. Our rationale was that if our assessment successfully measured participants' orthographic processing ability, then the assessment should be able to predict their spelling acquisition. Most importantly, if our assessment successfully distinguished orthographic processing from word recognition, our orthographic processing assessment would be able to account for additional variance beyond word recognition ability.

Moreover, we also examined internal structure and reviewed the potential consequences of using this instrument, as they are critical for examining the validity of an

instrument (American Education Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Kane, 2006, 2010, 2013; Wilson, 2005).

Data analysis

The computer program, ConQuest 2, was used to analyze the accuracy data. ConQuest 2 is used for fitting item response and latent regression models and allows for the examination of various properties of performance assessments (Wu, Adams, Wilson, & Haldane, 2007).

To examine the suitability of items, we followed the procedures suggested by Wilson (2005). A simple logistic Rasch model (Rasch, 1980) was used to fit the 140 cases. Both item and person fit statistics were examined. The point-biserial correlation and the mean location statistic were used to check that all items functioned in the direction intended. Item difficulty and the standard errors of measurement were evaluated. Differential item functioning between males and females as well as between Taiwanese and American students was investigated.

Results and discussions

We will first report descriptive statistics. Then, we will present item and person fit statistics to evaluate the suitability of the items and the sample. To further examine the success of the item design, the point-biserial correlation statistics and the mean location statistics are examined. A Wright map is presented to summarize the difficulty of items and the performance of participants. Standard errors and validity evidence are also reported.

Descriptive statistics

The mean, standard deviation, minimum, and maximum are summarized in Table 1. The letter and word recognition ability of Taiwanese participants, as measured by the Letter and Word Identification subtest of the Woodcock-Johnson III Test of Achievement (Woodcock et al., 2007), was lower than the American second-grade students in both statistical and practical significance, $t(101) = -14.07$, $p < .001$, Cohen's $d = 3.1$. Similarly, the Taiwanese participants performed lower on English oral vocabulary than the American second-grade students $t(101) = -36.33$, $p < .001$, Cohen's $d = 1.95$.

Item and person fit statistics

In the Rasch approach, fit statistics are used as a quality-control mechanism (Bond & Fox, 2015) to indicate the extent to which actual empirical responses differ from theoretical expectation. The weighted mean squares (infit mean squares) were used to examine the item and person fit statistics. The weighted mean square has an expected value of 1 (Linacre, 2017; Wright & Stone, 1979). A weighted mean squared statistic of 0.7 indicates less than 30% variance in an item's responses.

To compare infit statistics, both standardized (weighted t) and nonstandardized weighted mean squares (infit mean square) were used. Higher misfit values warrant

Table 1. Participants' performance on oral vocabulary, letter word identification, and orthographic processing.

Participants	*M*	*SD*	Min.	Max.
Taiwanese				
Oral vocabulary	28.76	13.40	13	64
Letter word identification	26.32	7.71	13	43
Orthographic processing	33.35	8.76	12	52
American				
Oral vocabulary	142.85	15.31	111	179
Letter word identification	49.24	7.45	36	65
Orthographic processing	36.30	9.89	13	54

further examination of items or persons. The standardized weighted mean squared values that fall outside the range of −1.96 to 1.96 are considered worth investigating. Lower than −1.96 fit values mean less randomness of responses than expected and higher than +1.96 mean more randomness than expected.

It is worth noting that, similar to an effect size, there is no absolute limit for a "good" weighted mean square value (Wilson, 2005). Adams and Khoo (1996) suggested a rule for determining misfit: unstandardized weighted mean squares larger than 1.33 or less than 0.75. Another commonly used rule is that those items with unstandardized weighted mean squares larger than 1.5 are a misfit (Boone, Staver, & Yale, 2014; Leung, Silverman, Nandakumar, Qian, & Hines, 2001; Linacre, 2017). An item is considered a misfit if it is out of range of both the weighted mean square statistic and the standardized statistic.

On this assessment, item 13 (*tplkmm*) had a standardized weighted mean square of 3 and an unstandardized weighted mean square of 1.20, representing 20% more variance than expected. This was the only item that was over the +1.96 limit for the standardized statistic. The point-biserial correlation of Item 13 was .17. Item 13 was an access task for which participants chose the letter string that was briefly displayed previously. Mean location analysis of Item 13 indicates that it was more difficult to correctly endorse this item than to provide an inaccurate response. Further inspection of the data logs for Item 13 did not reveal problematic issues, with the exception of one respondent who selected answer option "6" (possibly by accident) when only four options were presented. Thus this item was retained in the analyses.

One participant demonstrated misfit. Participant 131 had a standardized weighted mean square of 3.06 and a mean square statistic of 2.82. Participant 131 scored 59 out of a total possible of 60. Closer inspection of the kidmap revealed that this participant answered all items correctly except Item 29. Item 29 had an item difficulty of −1.02, whereas Participant 131 had an ability of 4.52. Thus, the misfit likely occurred because, based on his or her performance, Participant 131 should have had the ability to answer Item 29 correctly. We considered this response to be a result of random measurement error, and thus included this participant in our data analysis.

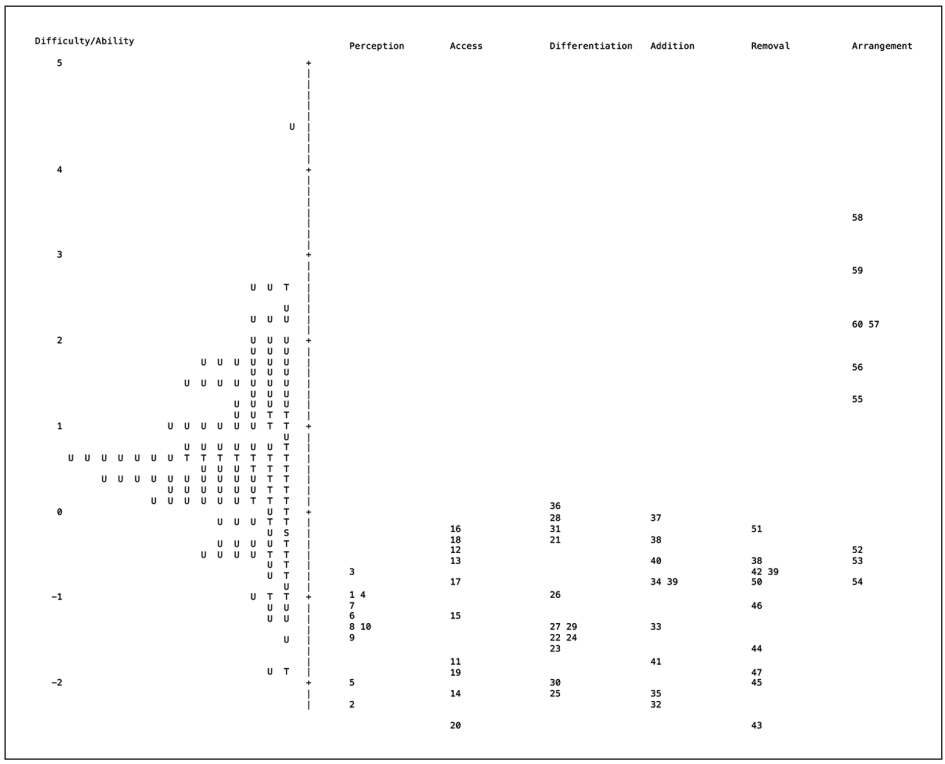

Figure 1. Wright map.

Precision

Person separation reliability indicates how well the items are able to separate the respondents' ability (Wright & Stone, 1979). Person separation reliability was .91 for this assessment, suggesting that participants' abilities were sufficiently well measured by the items. This indicates that it is appropriate to use the orthographic processing asessment to measure participants' ability. Cronbach's alpha for the EFL only sample and native speaker sample was .87 and .89, respectively, indicating that this test works equally well for less and more proficient readers. Standard errors of the person estimates ranged from .28 to .92. The participant with the highest ability had the largest standard error for person estimates, .92. The average standard error of the person estimates was .32. All of the items' estimated difficulties had a standard error of .30 or less, except Item 58 [*create*] which had a standard error of .42.

Wright map

The Wright map (Figure 1) summarizes item difficulty and person ability. The letters T and U represent participants from different geographical regions. T represents participants from Taiwan and U represents participants from the USA. Each number represents

one item: numbers 1 to 10 represent the perception task items; numbers 11 to 20 represent the access task; numbers 20 to 31 represent the differentiation tasks; numbers 32 to 51 represent the correction tasks (including addition and removal); and numbers 52 to 60 represent the arrangement tasks. No floor or ceiling effects were found, suggesting that this assessment was developmentally appropriate for these participants.

Participants' ability is concentrated at the lower end with a long upper tail. A participant from the USA had the highest ability, +4.52 logits. This participant also had the highest standard error for the person estimate, 0.92. Item 58 [*create*] had the highest item difficulty, 3.47 logits, whereas Item 20 [*birct*] had the lowest item difficulty, −2.20 logits. There is a gap between Item 55 (+ 0.50 logit) and Item 36 (+1.5 logits), suggesting the need to create more items with difficulties that fall between +0.50 logits and +1.5 logits.

All items that had difficulties higher than logit +2 were from the arrangement portion of the assessment. This pattern suggests that arrangement task items were more challenging than the correction task items, which was hypothesized a priori. It was also hypothesized that the arrangement task would be more challenging than the correction task because it required participants to produce an answer. In contrast, the correction task required participants to select an answer from a list of options. We did not hypothesize a difficulty ordering among the other four tasks for the following reasons: (a) the perception task had a response time limitation, and thus it cannot be readily copared with the other tasks which do not feature a response time limitation; (b) the stimuli for the access task was time limited, so it cannot be readily compared with other tasks that did not have a stimuli time limitation; and (c) the differentiation task used pesudowords as stimuli, so it should not be compared with the correction task in which real word stimuli was used.

Validity

Evidence based on internal structure. The Standards for Educational Psychological Testing (American Education Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) suggest using internal structure as one key component of validity evidence. Internal structure refers to "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (p. 13). We, therefore, tested whether a respondent with higher ability, in general, scored higher on each item. We evaluated the mean location and point-biserial correlation of each item. The point-biserial correlation of each item varied from +.17 to +. 63, suggesting that each item was positively correlated with a respondent's ability. Mean location analysis of each item indicated that it was more difficult to endorse an item correctly than to provide an inaccurate response, providing evidence of internal structural validity.

Additionally, differential item functioning was used to examine internal structure. Two types of differential item functioning analyses, that is, gender and country of residence, were tested. We used the rule suggested by Longford, Holland, and Thayer (1993) to identify which items have a larger than 0.426 logits difference, because a logit difference value less than 0.426 is negligible. In the gender analysis, among the 60 items, Item 16, Item 32, and Item 56 had logit difference values larger than 0.426, and these items

were more difficult for males. We were not able to ascertain a possible explanation for this difference and hence have not revised the items but have instead deleted them.

We also examined DIF between students in Taiwan and the USA. For items that had a larger than 0.426 logit difference, we found a unique and interesting pattern. All perception items were somewhat easier for the Taiwanese participants, as the difficulty estimates for all perception items (Item 1 to Item 10) were lower for the Taiwanese groups. This might be owing to the fact that the Taiwanese participants were more familiar with various symbol systems. In addition to learning the characters and logograms of Chinese, these students also learn the alphabetic system of English and Arabic numerals. An alternative explanation is that the Taiwanese participants were more skillful and more experienced in rapid counting because the Taiwanese participants were older than the American participants.

In contrast, the differentiation items (except Item 21) were somewhat easier for the American participants. This might be owing to the fact that the Taiwanese participants had limited exposure to English words; thus they may have incorrectly assumed that the pseudowords were new English words that they had not yet learned. In contrast, American participants had abundant exposure to English words, which may have helped them develop sensitivity to differentiate permissible and impermissible letter combinations. This phenomenon might suggest that, unlike phonological skills, orthographic skills are language-specific and do not transfer across languages (Abu Rabia, 2001; Arab-Moghaddam & Senechal, 2001; Wang, Park, & Lee, 2006; Wang, Perfetti, & Liu, 2006). The nontransferable nature of orthographic processing is especially evident when a learner's L1 and L2 languages have different scripts, such as English and Chinese (Commissaire, Duncan, & Casalis, 2011). Therefore, the Taiwanese participants could not take advantage of knowledge of orthographic patterns in their first language to complete the differentiation items. The lack of exposure to English words and the lack of support from their L1 may have caused the differentiation items to be more difficult for the Taiwanese participants. As this pattern is owing to the background of the participants, we decided to keep these items, but also hope that future research will examine the potential multidimensionality of the instrument.

Evidence based on relation to other variables (criterion validity). We provided two pieces of relational evidence between our assessment and other variables, including zero-order correlation and predictive and partial correlation.

Zero-order correlation. The orthographic processing assessment correlated with both oral vocabulary ($r = .39$) and word recognition ability, but the strength of the correlation was stronger for word recognition ability ($r = .59$). The relatively strong correlation between orthographic processing and word recognition suggests that our assessment successfully targets orthographic processing skills because it has been widely proposed that orthographic processing is a critical skill for word recognition (e.g., Cunningham, Perry, & Stanovich, 2001; Stanovich & West, 1989). However, it should also be noted that the moderate correlation ($r = .59$) between the orthographic processing assessment and word recognition ability suggests that our tool is not merely a measure of word recognition because if this were the case, the correlation would be stronger (i.e., $r \geqslant .80$). The weaker correlation between orthographic processing and oral vocabulary was expected because oral vocabulary does not involve the processing of letters or words, the skills assessed via this measure.

Table 2. Hierarchical multiple regression analyses predicting participants' spelling acquisition.

Predictors	Immediate spelling task		Delayed spelling task	
	ΔR^2	b	ΔR^2	b
Step 1				
Oral vocabulary	.22***	4.20***	.28***	4.85***
Step 2				
Oral vocabulary	.42***	1.24	.40***	1.83
Word recognition		6.50***		6.64***
Step 3				
Oral vocabulary	.11***	0.66	.9***	1.31
Word recognition		4.54***		4.87***
Orthographic processing		3.76***		3.39***
Total R^2	.75***		.77***	

Note: All predictors were transformed into z-scores. One participant was absent in the delayed posttest session. In total, there are 35 items on the spelling posttest.
b = standardized regression coefficient.
***$p < .001$. **$p < .01$.

Predictive and partial correlation. We also conducted a hierarchical regression analysis to examine the relationship between orthographic processing ability and participants' spelling acquisition (see Table 2). Hierarchical multiple regression analyses indicated that our orthographic processing assessment was a significant predictor of spelling acquisition after accounting for oral vocabulary and word recognition. After controlling for oral vocabulary and word recognition, orthographic processing explained an additional 11% of variance in participants' performance on an immediate spelling posttest and an additional 9% variance in participants' performance on a delayed spelling posttest. Holding oral vocabulary and word recognition ability constant, for every one z-score increase in orthographic processing, on average, participants' performance on the immediate spelling posttest was estimated to increase by 3.76; and participants' performance on the delayed spelling task was estimated to increase by 3.39. This suggests that for every one z-score increase in orthographic processing, on average, participants spelled three more words correctly. We consider this compelling evidence of the validity of our tool.

Consequential validity. Kane (2006, 2013) has advocated for the importance of specifying consequences in validity research. We, therefore, include here a discussion of the consequence of using this assessment, despite the fact that some researchers question the suitability of consequential validity in the validation process (Chalhoub-Deville, 2016).

The use of this assessment in literacy research will help educators and researchers to identify learners who may have difficulty in orthographic processing and to inform the design of suitable instructional support. By analyzing participants' responses to our assessment, we can pinpoint potential problems that hinder literacy development. For example, if students struggle with perception tasks, it suggests they are not yet able to recognize letters automatically and/or accurately. Thus, instruction should focus on

teaching English letters and on helping students to recognize letters automatically. If students struggle with the differentiation task, it suggests that they are not familiar with English spelling patterns. Thus, instruction should focus on teaching spelling patterns.

Conclusion

Currently, studies documenting the development of orthographic processing skills are rare, in part, because of the lack of appropriate assessments designed to assess pure orthographic processing. In order to bridge this gap in the research, it is necessary to design more accurate and comprehensive probes of orthographic processing (Conrad, Harris, & Williams, 2012). Our assessment has bridged this gap by including multiple tasks to assess the different aspects of processing by using both real words and pseudowords, resulting in the first comprehensive assessment of orthographic processing.

We examined the limitations of previous assessments of orthographic processing and designed items to overcome those limitations. For our novel assessment of orthographic processing, we used item fit statistics, mean location statistics, and point-biserial correlation to examine the suitability of test items. We also used differential item functioning to evaluate the performance of students in Taiwan as compared to those in the USA and the performance of female versus male students. As further validity evidence, we also examined the correlation of this assessment with external variables and the role of orthographic processing when making orthographic analogies during spelling acquisition. Our initial findings suggest that our assessment is a suitable and comprehensive tool for assessing an individual's orthographic processing ability.

Nevertheless, our initial evidence is limited in terms of our sampling method and a unidimensional assumption of the construct. Moreover, our results of DIF analyses are inconclusive owing to our small sample size. Future larger-scale research is needed to test the multidimensionality of the assessments to provide more conclusive evidence.

Another limitation is the length of the assessment. Because we wanted the assessment to be able to be completed within 30 minutes, we only included about 10 items per task. Future studies can create more items and use item difficulty and item discrimination statistics to build an item bank for orthographic processing.

Owing to the limitations of previous orthographic processing assessments, empirical research in the development of orthographic processing is not as comprehensive as phonological processing. We hope that our orthographic processing assessment will invite more research on orthographic processing, which will advance our understanding of the developmental trajectory of orthographic processing skills and yield a better understanding of its role in literacy development.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Yi-Jui Iva Chen https://orcid.org/0000-0001-9845-5764

Supplemental material

Supplemental material for this article is available online.

Note

1. Orthographic representations refer to English letters. Orthographic patterns refer to American English spelling rules.

References

Abu-Rabia, S. (2001). Testing the interdependence hypothesis among native adult bilingual Russian–English students. *Journal of Psycholinguistic Research*, *30*, 437–455. https://doi.org/10.1023/A:1010425825251

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. MIT Press.

Adams, R. J., & Khoo, S. T. (1996). *Quest*. Australian Council for Educational Research.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *The standards for educational and psychological testing* (2nd ed.). American Educational Research Association.

Apel, K. (2009). The acquisition of mental orthographic representations for reading and spelling development. *Communication Disorders Quarterly*, *31*(1), 42–52. https://doi.org/10.1177/1525740108325553

Apel, K. (2011). What is orthographic knowledge? *Language, Speech, and Hearing Services in Schools*, *42*(4), 592–603. https://doi.org/10.1044/0161-1461(2011/10-0085)

Apel, K., Thomas-Tate, S., Wilson-Fowler, E. B., & Brimo, D. (2011). Acquisition of initial mental graphemic representations by children at risk for literacy development. *Applied Psycholinguistics*, *33*(2), 365–391. https://doi.org/10.1017/S0142716411000403

Arab-Moghaddam, N., & Sénéchal, M. (2001). Orthographic and phonological processing skills in reading and spelling in Persian/English bilinguals. *International Journal of Behavioral Development*, *25*, 140–147. https://doi.org/10.1080/01650250042000320

Berninger, V. W., Cartwright, A. C., Yates, C. M., Swanson, H. L., & Abbott, R. D. (1994). Developmental skills related to writing and reading acquisition in the intermediate grades. *Reading and Writing*, *6*, 161–196. https://doi.org/10.1007/BF01026911

Berninger, V. W., Yates, C., & Lester, K. (1991). Multiple orthographic codes in reading and writing acquisition. *Reading and Writing*, *3*, 115–149. https://doi.org/10.1007/BF00420030

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer.

Burt, J. S. (2006). What is orthographic processing skill and how does it relate to word identification in reading? *Journal of Research in Reading*, *29*, 400–417. https://doi.org/10.1111/j.1467-9817.2006.00315.x

Carroll, D. W. (1999). *Psychology of language*. Brooks/Cole.

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

Castles, A., & Nation, K. (2008). Learning to be a good orthographic reader. *Journal of Research in Reading*, *31*, 1–7. https://doi.org/10.1207/s1532799xssr0902_4

Chalmers, K. A., & Burt, J. S. (2008). Phonological and semantic information in adults' orthographic learning. *Acta Psychologica*, *128*, 162–175. https://doi.org/10.1016/j.actpsy.2007.12.003

Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing*, *33*, 453–472. https://doi.org/10.1177/0265532215593312

Chen, Y.-J. I., Cunningham, A. E., Rabe-Hesketh, S., Hinshaw, S. P., & Irey, R. C. (2019). The effect of orthographic neighbors on second-grade students' spelling acquisition. *Reading Research Quarterly*. Advance online publication. https://doi.org/10.1002/rrq.291

Coltheart, M. (2005). *Modeling reading: The dual-route approach*. Blackwell.

Commissaire, E., Duncan, L. G., & Casalis, S. (2011). Cross-language transfer of orthographic processing skills: A study of French children who learn English at school. *Journal of Research in Reading*, *34*, 59–76. https://doi.org/10.1111/j.1467-9817.2010.01473.x

Conrad, N. J., Harris, N., & Williams, J. (2012). Individual differences in children's literacy development: The contribution of orthographic knowledge. *Reading and Writing*, *26*(8), 1223–1239. https://doi.org/10.1007/s11145-012-9415-2

Cummings, D. W. (1988). *American English spelling*. Baltimore, MD: The Johns Hopkins Press.

Cunningham, A. E. (2006). Accounting for children's orthographic learning while reading text: Do children self-teach? *Journal of Experimental Child Psychology*, *95*, 56–77. https://doi.org/10.1016/j.jecp.2006.03.008

Cunningham, A. E., Nathan, R. G., & Raher, K. S. (2011). Orthographic processing in models of word recognition. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research IV* (pp. 259–285). Routledge.

Cunningham, A. E., Perry, K. E., & Stanovich, K. E. (2001). Converging evidence for the concept of orthographic processing. *Reading and Writing: An Interdisciplinary Journal*, *14*, 549–568. https://doi.org/10.1023/A:1011100226798

Cunningham, A. E., & Stanovich, K. E. (1990). Assessing print exposure and orthographic processing skill in children: A quick measure of reading experience. *Journal of Educational Psychology*, *82*, 733–740. https://doi.org/10.1037/0022-0663.82.4.733

Deacon, S. H., Commissaire, E., Chen, X., & Pasquarella, A. (2012). Learning about print: The development of orthographic processing and its relationship to word reading in first grade children in French immersion. *Reading and Writing*, *26*, 1087–1109. https://doi.org/10.1007/s11145-012-9407-2

Dunn, L. M., & Dunn, D. M. (2012). *Peabody Picture Vocabulary Test, (PPVT-4)*. Pearson Education.

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

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

Hagiliassis, N., Pratt, C., & Johnston, M. (2006). Orthographic and phonological processing in reading. *Reading and Writing*, *19*, 235–263. https://doi.org/10.10007/s11145-005-4123-9

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*(3), 491–528. https://doi.org/10.1038/301419a0

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*, 177–182. https://doi.org/10.1177/0265532209349467

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73. https://doi.org/10.1111/jedm.12000

Kaufman, A. S., & Kaufman, N. L. (2014). *Kaufman Test of Educational Achievement–Third Edition (KTEA-3)*. Pearson.

Leung, C. B., Silverman, R., Nandakumar, R., Qian, X., & Hines, S. (2011). A comparison of difficulty levels of vocabulary in first grade basal readers for preschool dual language learners and monolingual English learners. *American Educational Research Journal*, *48*(2), 421–461. https://doi.org/10.3102/0002831210382890

Linacre, J. M. (2017). *Winsteps® Rasch measurement computer program: User's guide*. https://www.winsteps.com

Liu, X., Marchis, L., DeBiase, E., Breaux, K. C., Courville, T., & Pan, X., et al. (2017). Do cognitive patterns of strengths and weaknesses differentially predict errors on reading, writing, and spelling? *Journal of Psychoeducational Assessment*, *35*(1–2), 186–205. https://doi.org/10.1177/0734282916668996

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds), *Differential item functioning* (pp. 67–113). Lawrence Erlbaum.

Mayer, R. E. (1996). Learners as information processors: Legacies and limitations of educational psychology's second metaphor. *Educational Psychologist*, *31*, 151–161.

McClung, N. A., O'Donnell, C. R., & Cunningham, A. E. (2012). Orthographic learning and the development of visual word recognition. In *Visual word recognition* (pp. 173–195). Psychology Press.

McGrew, K. S., Schrank, F. A., & Woodcock, R. W. (2007). *Technical manual: Woodcock-Johnson III normative update*. Riverside.

Olson, R., Wise, B., Conners, F., Rack, J., & Fulker, D. (1989). Specific deficits in component reading and language skills: Genetic and environmental influences. *Journal of Learning Disabilities*, *22*, 339–348. https://doi.org/10.1177/002221948902200604

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383. https://doi.org/10.1080/10888430701530730

Perfetti, C. A., Landi, N., & Oakhill, J. V. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Blackwell.

Psychology Software Tools, Inc. [E-Prime 2.0]. (2012). https://www.pstnet.com/

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). University of Chicago Press.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568. https://doi.org/10.1037/0096-1523.13.1.14

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquistion. *Cognition*, *55*, 151–218. https://doi.org/10.1016/0010-0277(94)00645-2

Share, D. L. (1999). Phonological recoding and orthographic learning: A direct test of the Self-Teaching Hypothesis. *Journal of Experimental Child Psychology*, *72*, 95–129. https://doi.org/10.1006/jecp.1998.2481

Share, D. L. (2004). Orthographic learning at a glance: On the time course and developmental onset of self-teaching. *Journal of Experimental Child Psychology*, *87*, 267–298. https://doi.org/10.1016/j.jecp.2004.01.001

Stanovich, K. E., & Siegel, L. S. (1994). Phenotypic performance profile of children with reading disabilities: A regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology*, *86*, 24–53. https://doi.org/10.1037/0022-0663.86.1.24

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, *24*, 402–433.

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

Tsung, L., Zhang, L., & Hau, K. T. (2017). Contribution of working memory, orthographic and sentential processing to Chinese text comprehension by Tibetan and Yi students. *The Reading Matrix*, *17*, 16–39. https://pdfs.semanticscholar.org/541b/7b6690023ea489b9f55639286230bbe6112f.pdf

Wagner, R. K., & Barker, T. A. (1994). The development of orthographic processing ability. In V. W. Berninger (Ed.), *The varieties of orthographic knowledge, 1: Theoretical and developmental issues*. (pp. 243–276). Kluwer.

Wang, M., Park, Y., & Lee, K. R. (2006). Korean–English biliteracy acquisition: Cross-language phonological and orthographic transfer. *Journal of Educational Psychology*, *98*, 148–158. https://doi.org/10.1037/0022-0663.98.1.148

Wang, M., Perfetti, C. A., & Liu, Y. (2005). Chinese–English biliteracy acquisition: Cross-language and writing system transfer. *Cognition*, *97*, 67–88. https://doi.org/10.1016/j.cognition.2004.10.001

Wilson, M. (2005). *Constructing measures: A item response modeling approach*. Routledge.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2007) *Woodcock-Johnson III tests of achievement*. Riverside.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Mesa.

Wu, M., Adams, R., Wilson, M., & Haldane, S. (2007). *ACER ConQuest version 2.0: Generalised item response modeling software* [Computer software and manual]. ACER Press.