More WARs: The development of the WARL and the WARN

Kevin Wheldall

Robyn Wheldall

The MultiLit Research Unit has developed a series of assessment tools – curriculum-based measures (CBM) – that can be used to monitor the ongoing progress of students learning to read. *In a previous issue of* Nomanis, we reported the development of the Wheldall Assessment of Reading Passages (or WARP), which can be used to assess the fluency with which students read passages of text. The WARP is suitable for use with students who are reading at the Year 2 to Year 5 level (*Wheldall & Madelaine, 2000*; 2006).

In this current article, we describe the development by the MultiLit Research Unit of two other curriculum-based measures of reading fluency that are suitable for use with younger children who are performing at Year 1 and 2 levels: the Wheldall Assessment of Reading Lists, or WARL (*Wheldall et al., 2015*), and the Wheldall Assessment of Reading Nonwords (WARN; *Wheldall et al., 2021*). It is very important to have CBMs that can track progress across the first two years of schooling while students are (ideally) learning to read via explicit phonics instruction, and to have an efficient way of identifying students who are not making typical progress in the early stages of learning to read. By administering a test that identifies struggling students effectively, as early in the process as possible, teachers may be able to address the needs of struggling students in a timely manner and also to monitor their progress. This will result in fewer students being left to struggle for longer than necessary (*Bell et al., 2020*).

There are relatively few tests that measure general reading progress satisfactorily in the early years and far fewer still that allow monitoring on a regular basis. The two CBM assessment tools to be discussed here focus on the reading of single words (the WARL), and the reading of nonwords (the WARN).

To be of any practical use, any test or measure must be both reliable and valid. The authors of the test must be able to provide empirical evidence for the validity and reliability of their test. By validity, we mean the degree to which a test measures what it is supposed to measure. One of the most common ways of verifying if a new test is valid is by correlating the scores on the new test with scores on older tests that have already been established as valid indicators of reading performance (criterion validity). By reliability, we mean that the instrument must be capable of delivering the same result consistently. The test should give the same (or a very similar) result when it is given to the same child on separate occasions, close together in time. For example, if Mark scores 43 on the test on Monday (assuming that he has not been practising in between), then he should get a very similar score to 43 on, say, Wednesday, if the test is reliable. We call this test-retest reliability.

Psychometric property	Tests used	Correlational results
Participants: N = 122 Year 1 students (Reynolds et al., 2009)		
Parallel forms reliability	15 individual WARL lists	All list intercorrelations: .80–.97 (most coefficients over .90)
Participants: N = 335 (162 Year 1; 173 Year 2) students, assessed in February/March and again in August (Reynolds et al., 2011).		
Parallel forms reliability	WARN Initial Assessment Lists (Lists A, B and C) on both testing occasions	WARL Initial Assessment Lists inter-correlations: .93–.96
Test-retest reliability	WARN Initial Assessment Lists (Lists A, B and C), tested in February/March and retested in August	List A test-retest: .82 List B test-retest: .84 List C test-retest: .86 Average test-retest: .86
Criterion validity	Average from WARN Initial Assessment Lists; Martin & Pratt Nonword Reading Test; Burt Word Reading Test; South Australian Spelling Test (SAST); Sutherland Phonological Awareness Test – Revised (SPAT-R); Wheldall Assessment of Reading Passages (WARP)	WARL and Martin & Pratt: .75 WARL and Burt: .87 WARL and SAST: .83 WARL and SPAT-R: .83 WARL and WARP: .91

Table 1. Technical data (reliability and validity) for the WARL. All correlations significant at p < .001

Similarly, if the test has two different forms, say Form A and Form B, then they should provide very similar results. We call this parallel forms reliability. The most common measure of reliability is the correlation coefficient between the scores of the test on the two occasions it is given, or between the two forms of the test when they are given to a group of children.

This article will describe the construction of the WARL and the WARN and provide data on reliability and validity for both tests. This article also provides references to research we have carried out for the purposes of providing benchmark guidelines for the WARL and WARN. These benchmarks are guides based on a small but reasonably representative sample of students. Students who perform below the score designating the 25th percentile (bottom quartile) may be considered to be 'struggling' or low-progress readers and in need of reading intervention support. The 40th percentile scores provide minimum goals for students to achieve before exiting an intervention, in that scores within the 40th and 60th percentile range may be considered to be within the average range for literacy performance for that point in the school year. We hope that these benchmarks will provide rough approximations to guide instructional decision-making. It should be noted, however, that these are not 'norms' in the strict sense of being based on large representative samples of students.

Another brick in the WARL

We would like to acknowledge, at the outset, the major contribution of Dr Meree Reynolds in the development of this measure as part of her doctoral studies.

The Wheldall Assessment of Reading Lists (WARL) originally consisted of fifteen word lists. To construct the lists of words for the WARL, we started with a database of the 200 most common high-frequency single words found in children's storybooks and reading schemes read by five- to-seven-year-old children (*Stuart et al., 2003*). These 200 words were arranged into 20 groups of 10 words, with the words with the highest frequency being used in the first group and so on. Five words were randomly selected from each of these 20 groups and presented on a stimulus sheet as a 100-word reading task. This procedure was repeated 15 times to produce 15 alternative forms of the curriculum-based measure, each comprising 100 words.

The 15 100-word lists created were administered to a sample of 112 Year 1 students, who read each list for one minute each. Descriptive statistics for the 15 WARL lists (see *Reynolds et al., 2009*) showed that the means and standard deviations of the word list measures were relatively similar. Two of the word lists were subsequently excluded by a process in which consideration was given to both outliers and intercorrelations.

Following the procedure used when developing the WARP (see *Wheldall & Wheldall, 2020*), a decision was made to select three word lists from the remaining 13 lists, to be designated as the Initial Assessment Reading Lists. They were selected on the basis that they had the most similar means and standard deviations for words read correctly per minute. In addition, they correlated very highly with each other. The set of three Initial Assessment Word Lists of the WARL was deemed to be appropriate for screening procedures, for placement of students at appropriate levels of support, for pre- and post-testing in research studies, and for program evaluation. The mean of performance on the three lists is taken as the most reliable index, expressed in terms of words read correctly per minute.

The 10 word lists that remained were designated for monitoring progress during an intervention. The lists were very similar to one another in relation to their means and standard deviations. They also correlated highly with each other and with the mean score of the three Initial Assessment Lists. We suggest that if two WARL lists are administered fortnightly and averaged, the data is likely to be more reliable, smoother and more even in increments, enabling easier interpretation. We have produced a designated order in which the Progress Monitoring Lists should be used. When used in this order, the mean of each two successive progress tests is very similar.

Reliability and validity data for the WARL are summarised in Table 1 above.

Psychometric property	Tests used	Correlational coefficients
Participants: N = 163 (85 Foundation*; 78 Year 1) students from two schools with NAPLAN Year 3 results that were similar to national average.		
Parallel forms reliability	WARN Initial Assessment Lists (Lists A, B and C) and 5 sets of Progress Monitoring Lists (Lists 1-10)	All list intercorrelations: .97–.98
Criterion validity	WARN Initial Assessment Lists and Progress Monitoring Lists; Martin & Pratt Nonword Reading Test; Wheldall Assessment of Reading Lists (WARL)	WARN and Martin & Pratt: .85–.86 WARN and WARL: .91–.92
Discrimination	WARN Initial Assessment Lists, from Foundation and Year 1	Scores doubled from first to second year of schooling, showing good discrimination
Participants: N = 194 (101 Foundation*; 93 Year 1) students from four schools with NAPLAN Year 3 results that were similar to national average.		
Test-retest reliability	WARN Initial Assessment Lists (Lists A, B and C), tested in Term 2 and retested in Term 4	Average test-retest: .86
Criterion validity	WARN Initial Assessment Lists; Martin & Pratt; WARL	WARN and Martin & Pratt: .90 WARN and WARL: .89

*Foundation: first year of formal schooling
NAPLAN: National Assessment Program – Literacy and Numeracy

Table 2. Technical data (reliability and validity) for the WARN. All correlations significant at $p < .001$

Benchmark values for the WARL were *subsequently calculated*, for the average and bottom quartile scores of students at the beginning and middle of Years 1 and 2. These may be used as a guide for classroom teachers regarding typical progress.

Be WARNed

Measures of phonological recoding (nonword reading) and measures of reading fluency for students in the first two years of schooling are uncommon. (See *Colenbrander et al., 2011* for a review of nonword tests.) The Martin and Pratt Nonword Reading Test (*Martin & Pratt, 2001*) measures nonword reading but is not timed and offers only two forms. The Test of Word Reading Efficiency 2 (TOWRE-2; *Torgeson et al., 2012*) includes nonword reading and is timed but, again, has only two forms available. The Year 1 Phonics Screening Check, introduced by the UK Department of Education and now used in several states in Australia (*Department of Education, Skills and Employment, 2020*) is a one-off test given at the end of Year 1 that includes a measure of nonword reading but is, again, not timed.

There are relatively few tests that measure general reading progress satisfactorily in the early years and far fewer still that allow monitoring on a regular basis.

The Wheldall Assessment of Reading Nonwords (or WARN) is a new curriculum-based measure of nonword reading developed by the MultiLit Research Unit (*Wheldall et al., 2021*). The measure is intended as a quick and simple test to measure progress in learning phonic decoding skills (phonological recoding) during the early stages of reading skill development, and to identify young struggling readers. The advantage of the WARN over existing measures of phonological recoding is that it comprises multiple parallel forms, thereby allowing for continual monitoring of individuals over time.

The WARN consists of 13 lists of 50 nonwords. Three of the lists are used as the Initial Assessment Lists, and the remaining 10 lists form five sets of two Progress Monitoring Lists, to be used fortnightly for the purpose of tracking progress. The Initial Assessment Lists can be used for screening or as a post-test measure following an intervention, either after two school terms or at other intervals.

Students read from each list for 30 seconds to determine the number of nonwords read accurately within that timeframe, and their performance over three lists (Initial Assessment Lists) or two lists (Progress Monitoring Lists) is averaged.

The WARN offers content validity, as the test stimuli align closely with the content sequence of InitiaLit Foundation (InitiaLit–F), an instructional program which adheres to best practice according to the available theory and research (MultiLit, 2017). Nonword stimuli on the WARN were constructed using phonemes taught in the InitiaLit–F program. The words in each list follow the sequence of the phonemes in the program, which in turn was based on the principles outlined by Carnine et al. (2006).

The InitiaLit–F instructional program (MultiLit, 2017), which is targeted towards beginning readers, comprises 11 succeeding levels (known as 'sets') of instruction in letter-sound correspondences as part of a systematic synthetic phonics program. For the purpose of constructing the WARN, Sets 1 and 2 were combined to form 10 'sets' in total. Ten nonwords were generated from each of the reduced sequence of sets, using the letter-sound correspondences taught at each successive set. The nonwords were three or four phonemes in length (CVC, CCVC or CVCC; C = consonant, V = vowel), and included digraphs (for example, 'fim', 'juck', 'nump', 'swong').

Each WARN list was created by randomly selecting five nonwords from the 10 nonwords constructed at each set, yielding a list of 50 nonwords presented on a stimulus sheet. This process of randomly selecting five words from 10 options in each set was repeated 15 times to generate 15 lists, each comprising 50 nonwords.

All lists were administered to a sample of students in Foundation (i.e., first year of schooling) and Year 1. Means and standard deviations for each measure were calculated and all measures were inter-correlated. As expected, all 15 nonword lists produced very similar means and standard deviations and were highly intercorrelated (r = .92–.96, p < .001).

From these 15 lists, the most similar 13 lists were chosen and allocated to one set of three lists and five sets of two lists; the former to serve as the Initial Assessment Lists and the latter to serve as the Progress Monitoring Lists. The averages of these six sets were analysed to confirm that they were highly intercorrelated (r = .97–.98, p < .001).

Reliability and validity data for the WARN are summarised in Table 2.

Benchmark values for the WARN were calculated for the average and bottom quartile scores for students in the first and second years of schooling, as a guide for classroom teachers regarding typical progress (Wheldall et al., 2021).

Conclusion

Curriculum-based measurement is a quick, reliable, valid and cost-effective method of tracking progress in reading. It provides valuable information which enables educators to monitor progress regularly and to make appropriate instructional decisions in order to maximise the reading progress of their students. The series of CBM instruments we have developed (collectively known as the WARs) provide an effective Australian solution to monitoring students' reading progress.

But what of the future? A problem upon which we are still working is the development of yet another WAR, the Wheldall Assessment of Reading Comprehension or WARC. This is proving more difficult, but we continue to experiment with a maze procedure, whereby students need to select the seventh words from a 200-word passage out of a list of four plausible alternatives. Watch this space!

Recommended reading on curriculum-based measurement

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*(2), 188–192.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABCs of CBM: A practical guide to curriculum-based measurement* (2nd ed.). The Guilford Press.

Madelaine, A., & Wheldall, K. (1999). Curriculum-based measurement of reading: A critical review. *International Journal of Disability, Development and Education, 46*(1), 71–85.

Madelaine, A., & Wheldall, K. (2004). Curriculum-based measurement of reading: Recent advances. *International Journal of Disability, Development and Education, 51*(1), 57–82.

Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *International Scholarly Research Notices*, 2013. doi:10.1155/2013/958530

Disclosure

Kevin and Robyn Wheldall are directors of MultiLit Pty Ltd, in which they have a financial interest. They receive a benefit from the activities of the company and the sale of its programs and products, including the measures mentioned in this article.

This article originally appeared in the Learning Difficulties Australia Bulletin.

Emeritus Professor Kevin Wheldall [@KevinWheldall on Twitter], AM, BA, PhD, C.Psychol, MAPS, FASSA, FBPsS, FCollP, FIARLD, FCEDP, served as Professor and Director of Macquarie University Special Education Centre (MUSEC) for over twenty years prior to his retirement in 2011. He is Chairman of MultiLit Pty Ltd and Director of the MultiLit Research Unit and is the author of over three hundred academic books, chapters and journal articles. In 1995, he established the MultiLit (Making Up Lost Time In Literacy) Initiative, to research and develop intensive literacy interventions. He is a Fellow of the Academy of Social Sciences in Australia and in 2011 was made a Member (AM) in the Order of Australia.

Dr Robyn Wheldall (formerly Beaman) [@RWheldall on Twitter], BA, PhD, was a Research Fellow at Macquarie University until her retirement in 2011 and now continues as an Honorary Fellow. She is a founding director of the University spin-off company MultiLit Pty Ltd and is the Deputy Director of the MultiLit Research Unit. She jointly authored 'An Evaluation of MultiLit' (2000) (commissioned by the Commonwealth Government) and has published numerous articles in peer reviewed journals. Robyn has extensive experience in the establishment and implementation of intensive literacy programs in community settings. In 2005 she was awarded a Macquarie University Community Outreach Award for her MultiLit work.