The development and use of the WARs

Nicola Bell, Robyn Wheldall, Alison Madelaine and Kevin Wheldall

The WARs (our affectionate nickname for the Wheldall Assessment of Reading Passages, Wheldall Assessment of Reading Lists and the Wheldall Assessment of Reading Nonwords) are a set of assessment measures that are used to quickly gauge primary school students' reading proficiency. In this article, we will detail the process of developing these instruments, as well as the rationales underpinning that process, the tests' psychometric qualities and the recommendations for using them in the classroom.

Development of the WARs

The Wheldall Assessment of Reading Passages (WARP) was the first to be published by MultiLit in 2013. Its development can be traced back to the mid-90s, when the MultiLit Research Unit's Director, Kevin Wheldall, wrote 21 passages, each 200 words and of roughly equal difficulty. Research was conducted to establish the five passages most highly correlated with one another (Wheldall & Madelaine, 1997). Based on those five passages, additional studies provided evidence for:

- using words correct per minute, as measured in the first minute of the student reading, rather than averaging over the entire passage (Wheldall & Madelaine, 1997)
- reliability and validity of the measure (see final section of this article for definitions of these psychometric qualities) (Madelaine & Wheldall, 1998, 2002a; Wheldall & Madelaine, 2000; Wheldall & Madelaine, 1997)
- WARP scores predicting reading ability better than teacher judgement (Madelaine & Wheldall, 2002c, 2005).

In the early 2000s, there was a transition from studying five passages to studying three. These three passages would later be known as the 'Initial Assessment' Passages, while another set of 10 passages would be known as the 'Progress Monitoring' Passages. As well as establishing reliability and validity for the Initial Assessment forms (Madelaine & Wheldall, 2002b), research in the lead-up to publication was devoted to establishing benchmarks that could identify students as either at risk or average (Madelaine & Wheldall, 2002a, 2002b).

The next WAR to be developed was the WARL, or the Wheldall Assessment of Reading Lists. The development process for the WARL was shorter, because having the WARP's structure and administration guidelines as a foundation meant there was only a little fine-tuning that took place before publication. Instead of passages, WARL stimuli comprise lists of isolated, high-frequency words. These were originally taken from a database of the

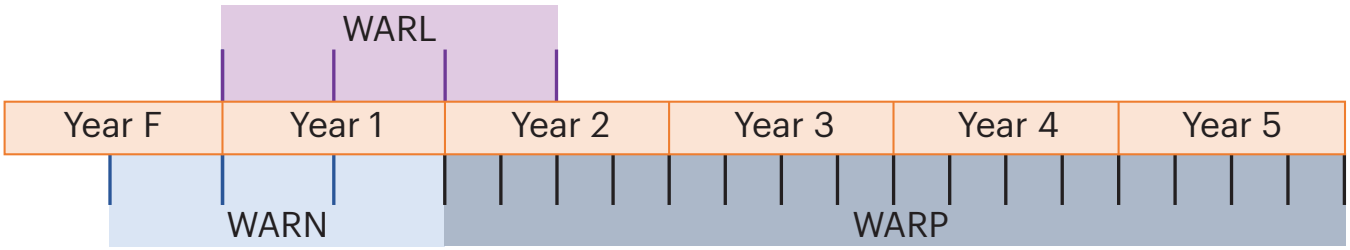

Figure 1. Timespans for average WARN, WARL and WARP benchmarks.

most frequently used words from books read by 5- to 7-year-olds.

The research underpinning the WARL's development established evidence for:

- using 100-word lists (Reynolds et al., 2009 [pilot study])
- having a 60-second duration for the test (Reynolds et al., 2009 [main study])
- 3 similar Initial Assessment Lists and 10 similar Progress Monitoring Lists (Reynolds et al., 2009 [main study])
- benchmarks for at-risk and average performance expectations (Reynolds et al., 2011)
- reliability and validity of the measure (Reynolds et al., 2009 [main study], 2011).

The most recent WAR to be published was the WARN, or the Wheldall Assessment of Reading Nonwords. The WARN stimuli were constructed using grapheme–phoneme correspondences taught in InitiaLit-F (MultiLit, 2017), although the stimuli aren't InitiaLit specific because they contain correspondences that are commonly taught in synthetic phonics approaches. In 2016, a proof-of-concept trial was conducted, followed by a more formal trial in 2017 and 2018. Results from these studies provided evidence for the same characteristics that are listed above for the WARL, except that having students read for just 30 seconds produced scores that were just as reliable as those for 60 seconds. This informed the decision to have the WARN be a 30-second measure, with just 50 nonwords to a page.

Rationale for conceptualisation of the WARs

All three WARs were designed to be curriculum-based measures (CBMs). This means they are intended to reflect students' skills in meeting the curriculum requirements. In terms of reading, CBMs do not need to strictly involve the same texts that are used in the classroom; it is enough that the measure represents the general reading curriculum. Hence, the overall rationale for developing the WARs was that they would give Australian teachers assessment tools that were quick and easy to administer but also effective in capturing students' reading proficiency.

Although they are all CBMs, the WARP, WARL and WARN look different because they are designed to reflect curriculum requirements of different year level ranges. To illustrate, Figure 1 shows a diagram of how the WARs fit together. The vertical lines mark the benchmarks that can be used for comparison against a student's score.

In the first couple of school years, a lot of the focus of reading instruction is on developing students' decoding skills and getting them to apply the letter–sound knowledge they have learned to sound out new words. The step beyond that is automaticity at a single-word level. At this point, it's hoped that the students have been applying their decoding skills and have started to build up a sight word vocabulary based on that self-teaching. Then, as we move up to Years 2 and 3, the focus turns to more passage-level reading and actually using texts to learn about other topics. Ultimately, the WARs reflect where students, very broadly speaking, are at in terms of their reading development, as well as what the reading curriculum is demanding of them at that point in time.

Rationale for format and structure of the WARs

The WARP is a classic example of an oral reading fluency measure. Oral reading fluency scores typically represent the number of words an examinee can accurately read aloud within one minute. The score therefore captures both accuracy and rate of reading aloud. Theoretically, it makes sense that both these factors contribute to overall reading comprehension. Readers must decode or recognise words to retrieve their meanings, and they must do this quickly enough to hold that meaning in mind while parsing the remainder of the sentence and passage. The relationship between oral reading fluency and reading comprehension has also been established empirically.

A reader's passage reading fluency depends on their automaticity of word identification. This is where the WARL – a measure of word identification fluency (WIF) – comes in. The WARL was intended to be more sensitive to changes in performance with younger readers and to be less daunting. It measures efficiency of word identification, so the factors of accuracy and rate are still what contribute to the score. In this case though, the score specifically reflects the reader's automaticity at a single-word level. Sight word retrieval efficiency is a key factor limiting reading comprehension. This is theorised in the Simple View of Reading model and, again, borne out in empirical research (Bell & Wheldall, 2022; García & Cain, 2014).

If we peel back another layer, we get down to nonword reading efficiency, which represents the accuracy and automaticity with which a reader can decode unfamiliar words. Skills in this area should theoretically feed into word identification efficiency via a self-teaching mechanism. Again, though, this relationship is not only theoretical because countless studies have shown that instruction focused on grapheme–phoneme relationships leads to improved word reading outcomes (Torgerson et al., 2006).

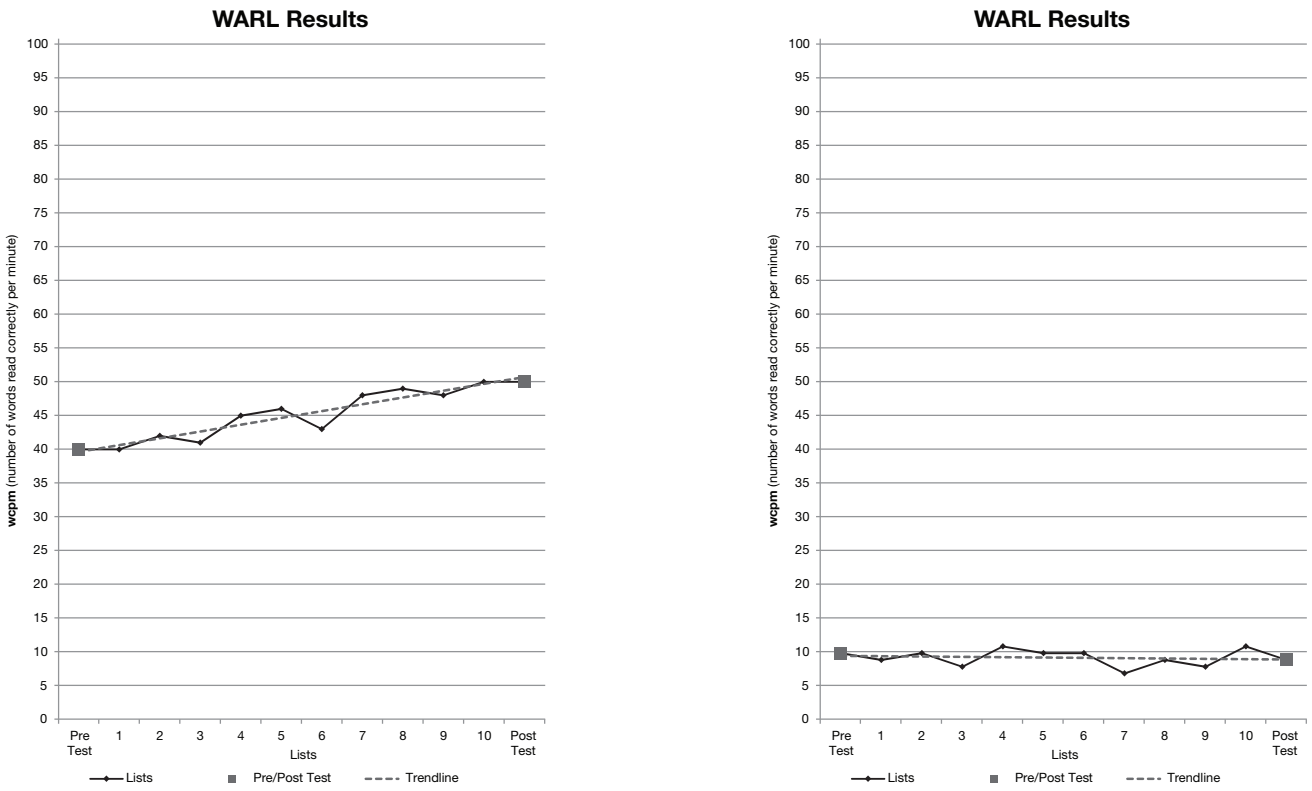

Figure 2. Progress monitoring graph showing increase vs. plateau in scores over time.

In summary, based on both theoretical and empirical grounds, it was decided that a timed oral reading task, with stimuli that aligned with grade-based expectations, would be the best format for a reading CBM.

Using the WARs in a Response to Intervention (or Instruction) approach

Within a Response to Intervention (RTI) approach, 'Tier 1' often refers to a whole-class teaching context, 'Tier 2' often refers to the small-group remedial support given to students with difficulties, and 'Tier 3' often refers to the individualised support given to students with significant difficulties.

Firstly, the Initial Assessment forms of the WARs can be used for screening in a Tier 1 setting, since the benchmarks allow users to identify students who may benefit from more targeted support in a Tier 2 setting. The Initial Assessment forms may also be administered at a whole-class level for long-term progress monitoring (e.g., 3–4 times per year). The purpose here would be to check that students are on track, and, again, to identify whether any students have fallen through the cracks and would benefit from Tier 2 intervention.

At a Tier 2 level, the Initial Assessment forms can be used as evidence in support of exiting a student from a program. Each measure's threshold scores for 'average' performance provide a goal that educators can aim for their students to reach through intervention. Once the students reach this goal, they may no longer be suitable for intervention and can move back into a Tier 1 teaching context.

Finally, at both Tiers 2 and 3, the Progress Monitoring forms for all WARs can be used to track students across shorter intervals (e.g., each fortnight). If they aren't responding to small-group instruction, they may need to move from Tier 2 to Tier 3. If they aren't responding to support delivered at an individual level, something more needs to be done. A comprehensive assessment of their language and cognitive abilities is warranted if that has not already been conducted.

Figure 2 illustrates why frequent progress monitoring can be useful for making instructional decisions. The first student is clearly responding positively to the intervention provided, since their scores are moving in the right direction towards average performance. On the other hand, the second student is not responding well. Assuming the intervention has a solid evidence base and is being delivered with fidelity, this would be considered a red flag, indicating that they may benefit from more individualised intervention and/or the additional support of a speech pathologist, specialist teacher or educational psychologist.

Strengths and limitations of the WARs

As outlined in the previous section, the WARs have multiple uses within an RTI context. They are quick and easy to administer, are sensitive to small improvements, allow for progress monitoring and have good reliability and validity (see Figure 3).

'Reliability' refers to the test's consistency across different testers, forms and testing times. A reliable test assesses what you want without capturing too much 'noise'. All three WARs have 'alternate forms' reliability at or above .9, which is excellent. 'Validity' refers to the test's ability to capture the specific skills of interest. We judge this by looking at how well a test correlates with other similar and dissimilar measures. As can be seen, the WARP is most strongly correlated with measures of passage reading accuracy and sight word reading, which is what you would expect of a test that theoretically aligns closely with these areas. This is a similar story for the WARL and WARN. In all, there is good evidence for the validity of each WAR.

As well as noting the strengths of the WARs, it's important to note their

Reliability (Alternate forms)

WARP	WARL	WARN
.97[a]	.90[b]	.94[c]

Validity

WARP	
WARL	.91[e]
NARA (passage) accuracy	.86[d]
Burt (sight word accuracy)	.83[d]
SAST (spelling)	.77[d]
Martin & Pratt (NW reading)	.59[d]
NARA Reading Comp	.55[d]
PPVT (vocabulary)	.33[d]

WARL	
TOWRE Sight Words	.92[b]
WARP	.91[e]
Burt (sight word accuracy)	.87[e]
WARN	.86[c]
SAST (spelling)	.83[e]
SPAT-R (PA)	.69[e]
TOWRE Phonemic Decoding	.76[b]
Martin & Pratt (NW reading)	.75[e]
PPVT (vocabulary)	.42[e]

WARN	
Martin & Pratt (NW reading)	.87[c]
WARL	.86[c]

[a]Madelaine & Wheldall (2002b)
[b]Reynolds et al. (2009)
[c]Wheldall et al. (2021)
[d]Wheldall et al. (in preparation)
[e]Reynolds et al. (2011)

Figure 3. Reliability and validity of the WARs.

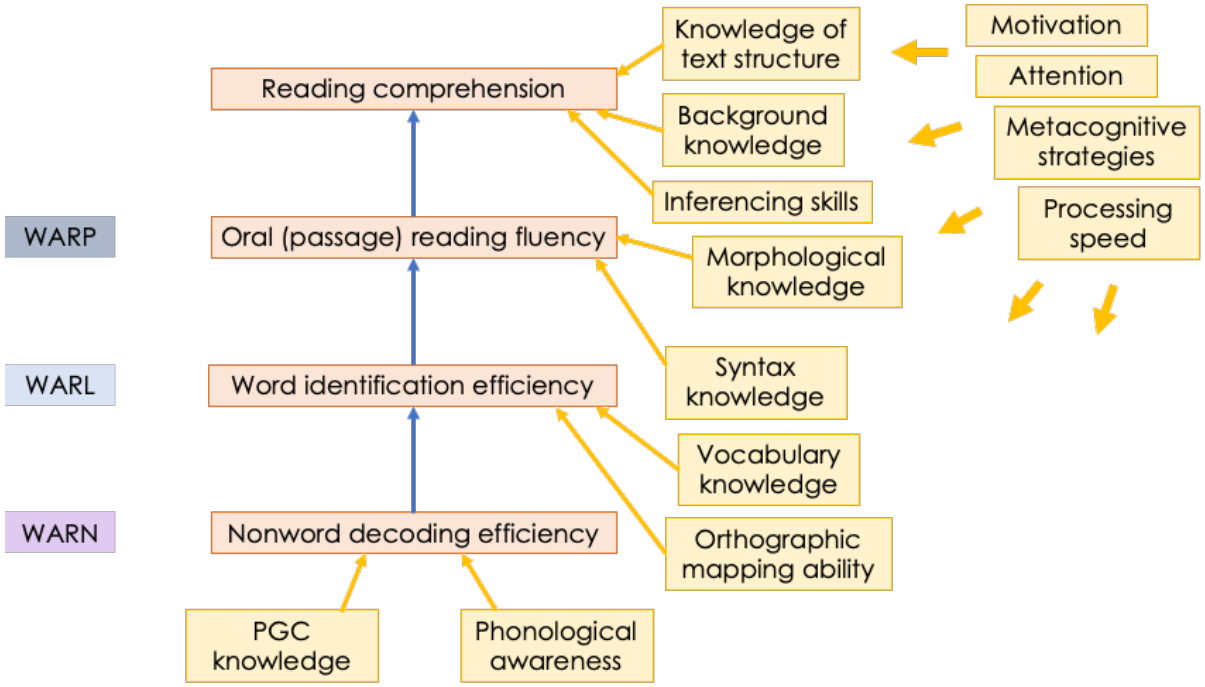

Figure 4. Factors contributing to reading comprehension.

limitations. They cannot and do not test all skills and factors that contribute to reading comprehension, as illustrated in Figure 4. These additional factors may be captured to varying degrees by the WARs, but further assessments would be required to draw any concrete inferences about their functioning.

The point is that the WARs are intended as quick and easy measures that index a student's developing reading proficiency. They provide useful information in that regard. However, they aren't intended to replace a comprehensive testing battery if that's what is considered necessary for a student to receive. This makes the WARs both limited and also fit for purpose.

Disclosure statement

Emeritus Professor Kevin Wheldall and Dr Robyn Wheldall are directors of MultiLit Pty Ltd and receive a benefit from the activities of the company and the sale of its programs and products, including the measures mentioned in this article. Dr Nicola Bell and Dr Alison Madelaine are paid employees of MultiLit Pty Ltd.

This article is an edited excerpt from a presentation delivered at Learning Difficulties Australia's 'Best Practice Using an RTI Framework' Online Conference.

The authors of this article (Nicola Bell [@NicolaBellSP], Robyn Wheldall [@RWheldall], Alison Madelaine [@alisonmadelaine] and Kevin Wheldall [@KevinWheldall]) are members of the MultiLit Research Unit.