When tests get old: A simple model for rehabilitation

Nicola Bell

Kevin Wheldall

What happens to tests that no longer produce accurate scores?

Background

Classroom-based assessments of reading have many purposes, one of which is to indicate to educators how a student's decoding skills compare with those of their peers. To gather such information, teachers and literacy specialists need tests that:

- are norm-referenced (i.e. they have norms that allow for a given score to be compared with a reference group of students in the same year level or age group)
- provide an accurate indication of ability (i.e. they measure the underlying construct of interest and provide a score that estimates the student's level of skill well)
- measure nonword reading performance (i.e. they contain items that must be decoded using knowledge of letter–sound correspondences).

At MultiLit, a test that we often use to assess nonword reading accuracy is the 'Martin and Pratt Nonword Reading Test'. In fact, the Martin and Pratt has been a staple within our program trial test batteries over many years, such that, even when it was no longer available for purchase, the MultiLit Research Unit (MRU) sought permission to continue using the materials.

In 2021, the MRU decided to embark on a 'check norming' study of the Martin and Pratt, since, at that point, 25 years had passed since the original test norms were collected. The aim was to confirm whether the assessment was valid and had standardised norms that accurately represented students' decoding skills. Specifically, the research questions of interest were:

1 How valid is the Martin and Pratt as a measure of nonword reading proficiency?

2 How well do Martin and Pratt standardised scores estimate primary school-aged students' nonword reading proficiency?

The two research questions sound alike but involve different methodologies and analyses.

The first question relates to *validity* – that is, the degree to which the test measures the underlying skill it is purported to measure. This is typically examined by analysing the correlations between the test of interest and other similar tests. The test may be said to have validity if the scores derived from it correlate strongly with other similar measures.

The second question relates to *norm representativeness*. Even if a test is valid, that doesn't mean its norms are up-to-date or representative. It may still measure what you want to measure but spit out a score that over- or

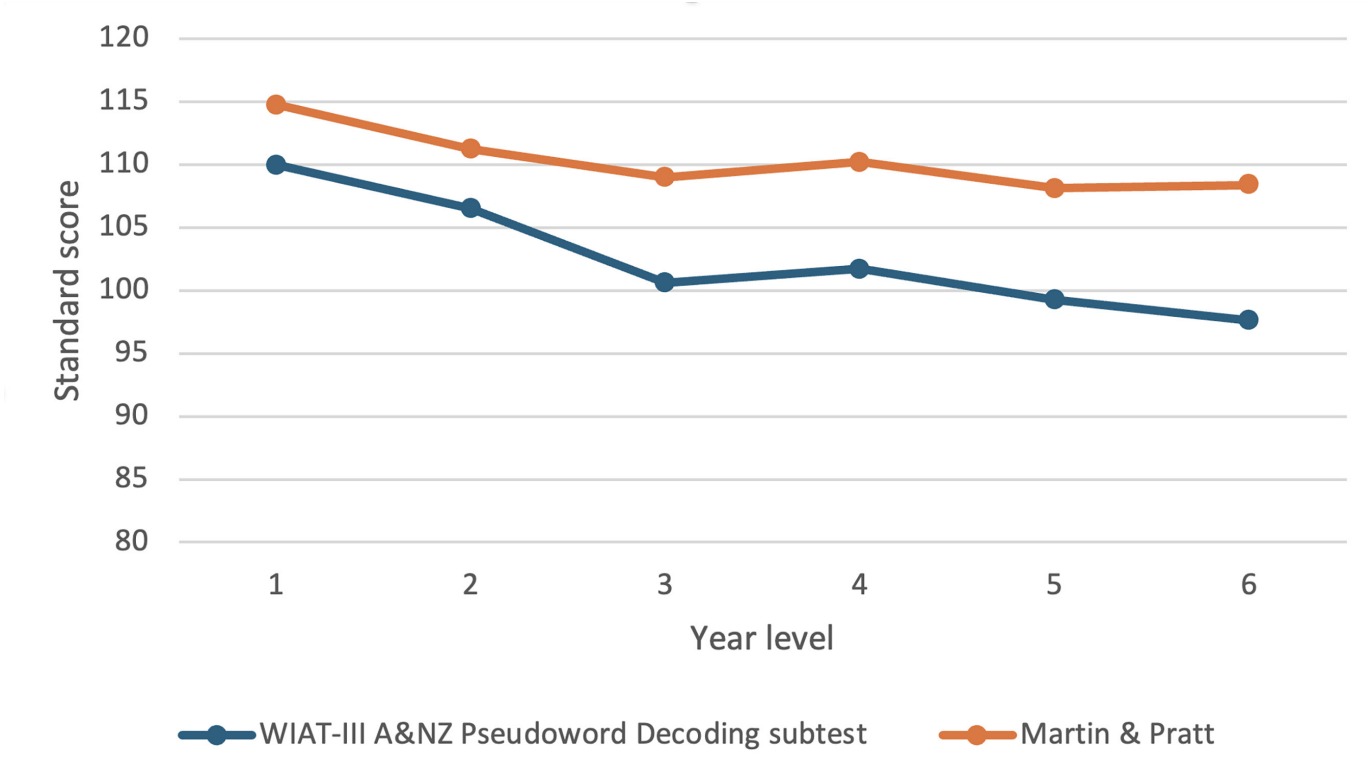

Figure 1. Mean standard scores achieved by each year level on the M&P and WIAT-III A&NZ PD.

underestimates the examinee's actual skill in that area.

Data collection

The process for conducting the study started with obtaining the prerequisite ethics and research approvals. Getting institutional ethics approval from Macquarie University was a straightforward enough process. However, we did not receive approval from the Queensland Department of Education, which meant we could not recruit any government schools in Queensland into the study (and we could not assess any students within the grounds of government schools that approached us). As such, we only approached independent schools in the greater Brisbane area with information about the study, and we were fortunate to have three sign on to participate. In terms of sample size, we got close to the anticipated number of students – 176 in total, with close to 30 in each year level. The student populations at all schools had average socio-educational backgrounds and produced approximately average NAPLAN Reading results.

In addition to the Martin and Pratt, students were assessed on two other nonword reading accuracy measures: the Castles and Coltheart 2 (CC2) and the Wechsler Individual Achievement Test 3rd ed Australian & New Zealand Pseudoword Decoding (WIAT-III A&NZ PD) subtest. Other aspects of reading proficiency were also assessed, such as fluency and comprehension. Each assessment session lasted approximately an hour. One key correspondent from each school kept our testing team informed about the best days and times for withdrawing students from their classes for testing.

Results from the study

The results from our analyses showed that Martin and Pratt scores correlated significantly with all other reading measures. Importantly, the Martin and Pratt was most strongly correlated with the other two measures of nonword reading accuracy: CC2 Nonwords (r = .92) and WIAT-III A&NZ PD subtest (r = .91).[1] These findings provide good evidence for the Martin and Pratt's validity. As such, the test's raw scores can still reliably be used to measure a student's progress.

However, the *standard* scores derived from the Martin and Pratt were consistently higher than those derived from the other two nonword reading accuracy assessments (on average by 7 standard score points). This means that, despite the test's solid design and observed validity, the standard score and age equivalent values that the norms

These findings provide good evidence for the Martin and Pratt's validity.

[1] Generally, r-values above .80 are considered to represent a 'strong' relationship.

allow users to compute, significantly over-estimate students' skills. Figure 1 shows the mean standard score for each of the Martin and Pratt and WIAT-III A&NZ PD across year levels. The gap between scores that *should* assess the same underlying skills is obvious.

Recalibration of Martin and Pratt norms

Having found that the test norms over-estimated students' skills, we decided to try and recalibrate them. This involved a novel analytical process, which was based on the assumption that the WIAT-III A&NZ PD standard score represented students' *actual* level of nonword reading proficiency. (The WIAT-III A&NZ was selected for this purpose because its normative data were collected quite recently from students across all Australian states.) Using a few different techniques, we sought to close the gap between those WIAT-III A&NZ PD scores and the Martin and Pratt scores.

The technique we landed on involved conducting a regression analysis with Martin and Pratt standard scores, WIAT-III A&NZ PD standard scores and age as variables. The resulting equation from that analysis was used to update all values in the original norms table to 'recalibrated' values.

There are a couple of important differences between our recalibrated norms and the original ones. Firstly, we only used Form A of the test, whereas the original manual contains norms for both Form A and Form B. Secondly, the recalibrated norms extend only to 11 years, 11 months, whereas the original norms extend all the way to 16 years, 11 months. We intentionally limited the age range of our sample in this way because we have doubts about the meaningfulness of assessing nonword reading in typically developing students beyond the primary years.

Final thoughts

The study described here provided the impetus to attempt a 'rehabilitation' of the Martin and Pratt. Happily, and with the generous support of the original test authors, Frances Martin and Chris Pratt, preparations are now underway to republish the test alongside the recalibrated norms.

But one question worth asking is: Why were the Martin and Pratt norms no longer accurate in the first place? Obviously, a considerable amount of time has passed since data for the original test norms were collected, but what exactly happened during that time to change students' nonword reading ability at a population level?

We think the most likely answer is that Australian teacher knowledge around the importance of phonics has increased. More broadly, reading skills in general – as measured in primary school-aged students – have improved, according to national and international testing.

Based on those observed improvements alone, test developers should think more about how norms can be updated in response to widespread shifts in instructional practices – particularly when those practices have such a direct relationship with the skills being assessed (e.g. phonics instruction and nonword reading proficiency; see article by Shanahan in this issue of *Nomanis*). Perhaps the recalibration undertaken as part of our study could be useful as a model for 'rehabilitating' other assessments that are outdated.

Test developers should think more about how norms can be updated in response to widespread shifts in instructional practices – particularly when those practices have such a direct relationship with the skills being assessed.

This article is an edited version of a presentation delivered at the DSF Language, Literacy & Learning 2024 conference.

Nicola Bell [@NicolaBellSP on X] is a Research Fellow in the MultiLit Research Unit.

Kevin Wheldall AM [@KevinWheldall on X] is Chairman and Director of MultiLit and Emeritus Professor of Macquarie University.