Benchmarking assessments and levelling should be consigned to history

Jennifer Buckingham

In Episode 5 of Emily Hanford's podcast *Sold a Story,* there is a discussion about a question that comes up often: **are benchmarking assessments and levelled texts scientific evidence-based approaches to reading instruction and intervention?**

The short answer is no. Benchmarking assessments are a form of assessment called an Informal Reading Inventory (IRI). They are not standardised and publishers do not always evaluate and report their validity and reliability, and those that do often have significant caveats (*Spector, 2005*; *Nilsson, 2008*; *Nilsson, 2013*). Reliability refers to the stability, or consistency, of test scores; validity refers to the test itself and how well the test measures what it claims to measure.

Studies from a team of researchers at the University of Minnesota, Minneapolis, have shown that the Fountas and Pinnell Benchmarking Assessment System is not a reliable measure of reading ability or reading progress. (Both papers summarised below can be accessed via Researchgate).

In *Parker et al. (2015)*, second and third grade students were given an *oral reading fluency* (ORF) assessment and an IRI (in this study, the Fountas and Pinnell Benchmark Assessment System) to compare the diagnostic accuracy of the two assessments for identifying students considered at risk for failing a district-wide reading assessment.

Findings included:

- ORF and IRI results were correlated
- "However, ORF demonstrated higher diagnostic accuracy for correctly identifying at-risk students and resulted in 80% correct classification compared to 54% for the reading inventory data".

A secondary question addressed in these studies is: **is assigning a book 'level' based on the results of benchmarking assessments a valid way to guide and build students' reading ability?**

Once again, the short answer is no. The 'text gradient' levelling system for books is also highly variable and unreliable.

In *Burns et al. (2015),* second and third grade students read for one minute from three levelled texts that corresponded to their instructional level as measured by an IRI assessment (Fountas and Pinnell Benchmark Assessment System), and the percentage of words read correctly was recorded (using a words correct per minute [wcpm] measure).

The Fountas and Pinnell Benchmark Assessment System assigns a 'letter level' that corresponds to a set of books at that level.

Findings included:

- "[T]he categorical scores (frustration, instructional, and independent) for the three readings agreed approximately 67% to 70% of the time, which resulted in a kappa estimate of less than .50. Kappa coefficients of .70 are considered strong indicators of agreement."
- "One quarter of the time, the students read 93% to 97% of the words correctly when reading the book that was rated at their instructional level, and students who were struggling readers frequently failed to read at least 93% of the words correctly when they were reading from a book designated by an IRI to provide an appropriate level of difficulty."

There is noticeable variation in books within a single level, and no quantifiable or codifiable gradation between levels in even one levelled book series. There is no consistency in the levels among different series of levelled books. Therefore, the 'level' of a book is almost meaningless. In sum, if the benchmarking assessments have a high margin of error, and the system of book levels is also imprecise, we can't have much confidence that either is a good indicator of a student's reading ability and they are therefore a poor basis for instruction.

Furthermore, some researchers (e.g., Tim Shanahan) question the whole premise of frustration/independent/instructional level as a useful method of text selection. So, we could think of benchmarking and levelling as a waste of time either way, whether it's a reliable system of text-student matching or not.

- Shanahan, T. (2014). Should we teach students at their reading levels? *Reading Today*, September/October 2014.
- Shanahan, T. (2020). Limiting students to books they can already read: Why it reduces their opportunity to learn. *American Educator*, Summer 2020.
- Shanahan, T. (2021). What does the Easter Bunny have in common with the independent reading level? *Shanahan on Literacy*, 13 February 2021.

A recent webinar by Tim Shanahan describes a number of studies showing that students had more growth in reading when they read books that were harder than their 'instructional level' (with some cautions and exceptions outlined below). There is evidence that in paired oral reading activities such as dyad reading, it is beneficial for both the lead (higher ability) and assisted (lower ability) student in the pair to read books that are much more difficult than their 'instructional level' (Trottier Brown et al., 2017). Concerns about systems of levelling and text-student matching have also been raised from an inquiry-based perspective (Hoffman, 2017).

This raises the obvious question: **are reading programs that use this benchmarking and levelled text system evidence-based and effective?**

No surprises that the answer is again, no.

Fountas and Pinnell's program is not the only reading program that uses IRI assessments and levelled texts. *PM Benchmark Literacy Assessment* also uses this system of administering IRI assessments and assigning text levels using the well-known PM levels of 1–30 (PM stands for Performance Measurement). Yet, based on the evidence above, it is hard to imagine how reading programs like these could be effective in improving students' reading.

In addition, reading programs that use levelled text are designed around the disproven and ineffective three-cueing system for reading.

An evaluation of *Fountas & Pinnell Classroom* (K-2 and 3-5) by EdReports found that it "does not meet expectations" in all grades because it does not include evidence-based approaches to reading instruction such as systematic and explicit phonics instruction, among other weaknesses.

In Grades K–2, for example:

- It takes an analytic approach to teaching phonics with no evidence-based scope and sequence; only 10 minutes of phonics in a session; phonics is not taught daily; there is

no decodable text.

- No sequence for high frequency words.

In Grades 3–5, for example:

- Text quality and complexity is not appropriate (ironic given the program is text-based)
- Insufficient time on vocabulary and grammar
- Limited word analysis (including phonics)
- Fluency is not part of core instruction
- Writing instruction is intermittent.

According to *Professor Mark Seidenberg*, "Fountas and Pinnell's approach to reading creates learning difficulties for which their curriculum then offers solutions." EdReport's evaluation of Lucy Calkin's *Units of Study* received equally poor ratings for Grades *K-2* and *3-5*.

The Fountas and Pinnell *Levelled Literacy Intervention* (LLI) program also uses the Benchmark Assessment System and *Text Level Gradient*. Two studies of LLI in K-2 that meet the *ESSA* evidence standards had an average effect size of +0.13 on reading outcomes, which is statistically significant but negligibly small. Effective reading interventions achieve effect sizes in the order of +0.39 (*Gersten et al., 2020*).

The big question, therefore, is: **what should be used instead of benchmarking and text levels?**

All students should receive systematic and explicit instruction in the five essential components of reading identified by scientific reading research in the first years of school. This is becoming more widely accepted but a lot of teachers are reluctant to give up benchmarking and levelled texts even if their system doesn't require them. One reason might be that it is a process and a system they are familiar with, and that parents are familiar with, even if they know it's imperfect. Another reason might be that they don't know what to do instead.

In terms of assessment, IRIs and their close cousin Running Records are not fit-for-purpose. They do not give teachers the depth of information they need to make instructional decisions because a) they do not test the reading sub-skills that have been shown to contribute to reading fluency and comprehension, and b) they are not constructed or validated in such a way that allows a student's reading to be compared to their peers or that allows their reading progress to be measured and evaluated against benchmarks for risk. For young readers, alternative assessments should include phonic decoding and oral reading fluency. See the *Primary Reading Pledge* for more details. For older readers, oral reading fluency is still a strong measure of reading progress and highly correlated with comprehension. Reading comprehension assessments are fallible but a well-constructed comprehension assessment that has clear objectives can provide useful information. The new Comprehension section on the *Five from Five website* will have more information on assessment.

In terms of text selection, students who are still learning to decode and read words with automaticity should be using decodable texts for oral reading practice. They should still have access to other books and be engaged in shared reading with a wide range of children's literature and non-fiction for language and comprehension development.

More research is required on text selection for older students without reading difficulties but there are a couple of general guides based on the extant evidence. When students are able to decode proficiently, their choice of texts for oral reading practice should not necessarily be limited to an 'instructional' or 'independent' reading level. (For fluency instruction, it's a different rule of thumb; a text that is too hard will not allow a specific focus on developing fluency). Allowing and encouraging students to read more challenging texts will expose them to more vocabulary and more complex sentence structures, but it is important that this is supported to ensure that they understand what they are reading so that they can learn and improve. Throwing students in the deep end without these supports might be counterproductive (*Amendum et al., 2016*).

It is impossible to explicitly teach all the vocabulary and knowledge that is valuable to students – most of what they learn will be through reading. The task of the teacher is to calibrate instruction and practice so students are reading to learn while they are learning to read and vice versa.

And finally: **what can be done with all the levelled books I have in my classroom or school?**

This has been addressed in previous Five from Five blogs (*here* and *here*) and *Reading Rockets* also has good advice. To summarise, the lowest levels of levelled book series, which are typically predictable texts, should not be given to beginning readers. They can be creatively re-purposed. Other levelled books can just be treated like any other book. Don't rely on the letter or number level of the book and take a more individualised approach to which books will provide a student with a sufficient level of challenge.

Dr Jennifer Buckingham [@buckingham_j on Twitter] is Director of Strategy and Senior Research Fellow at MultiLit.