The Observation Survey and the National Standards

The purpose of this work is to inform overall teacher judgement of student performance against the 'after one year' national standard in reading[1]. In particular it will help teachers who use the Observation Survey as part of the evidence informing their overall teacher judgement.

The work

A team of experienced teachers and literacy professional development facilitators worked through sets of anonymised Observation Survey records of student performance (called 'scripts' in what follows).

Initially, the 'after one year' standard was introduced. This was followed by a description of how data collected using the Observation Survey could reveal features of the standard such as the ability to read, respond to, and think critically about texts. Then, after a discussion of the definitions of the standard's four reporting bands ('well below', 'below', 'at' and 'above'[2]) and a practice attempt at rating a script, the experts made independent judgements on a pack of six scripts each. Altogether 30 scripts were rated against the standard, with each being rated by two separate judges.

The teachers and other judges made their decisions independently, so that a range of judgements for a given level of performance was captured. It means that for any one piece of evidence describing student performance, (such as the Observation Survey), only the likelihood of that piece of evidence being judged as 'well below', 'below', 'at' or 'above' the relevant national standard can be provided.

The number of scripts used in this work, (30), was modest and the work has yet to be replicated on a larger scale. The Ministry has done similar work for the PAT: Reading and STAR assessments. It intends to follow this with similar work for other assessment tools.

Results

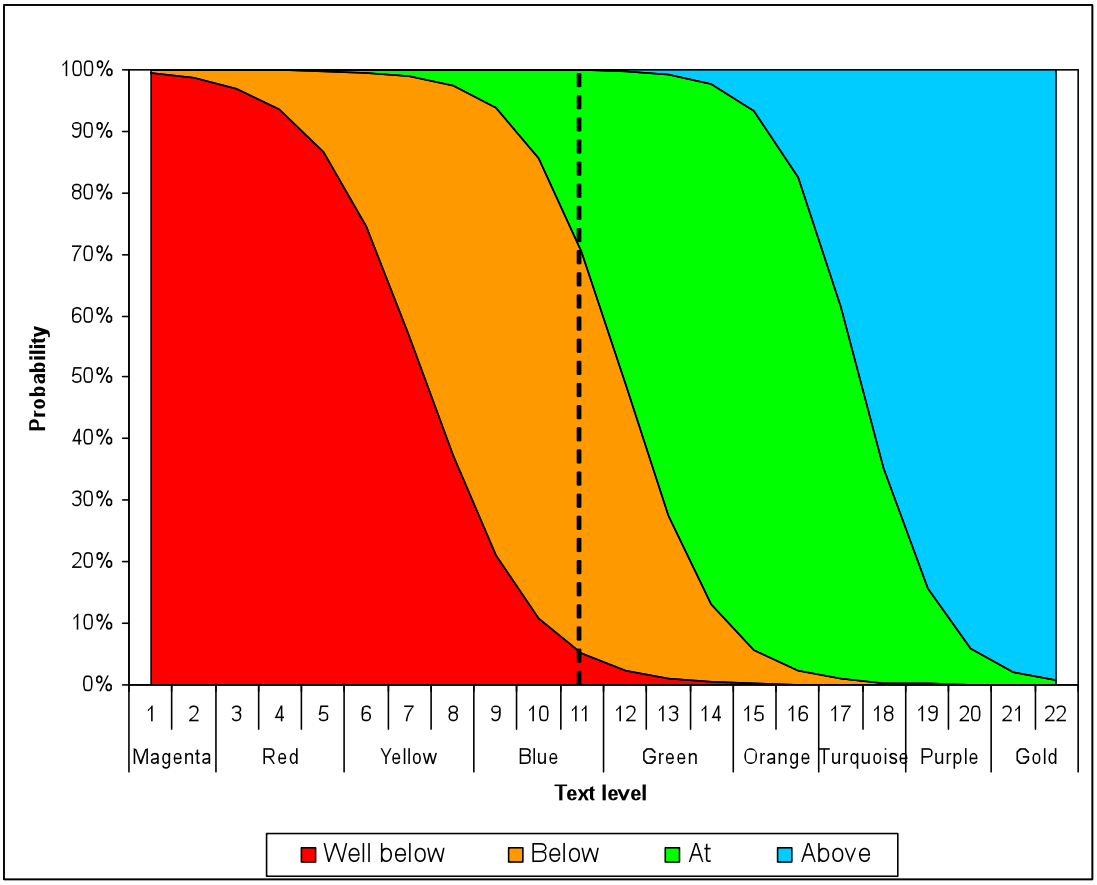
Student achievement in the Observation Survey is measured by the maximal text level (of at most three) in the Running Record for which the student had an error rate of 10% or less as well as by the stanine scores relating to the five 'test' elements of the assessment (letter identification, concepts about print, writing vocabulary, word reading, and hearing and recording). The current results describe how these measures align with the 'after one year' national standard in reading. The 'area graph' below shows the percentage of scripts in each of the four reporting bands against the reading text level (i.e. the highest level with an error rate of 10% or below). It was produced using a statistical modelling technique applied to the collected data.

---

[1] See http://nzcurriculum.tki.org.nz/National-Standards/Reading-and-writing-standards/The-standards/After-one-year

[2] See http://assessment.tki.org.nz/Effective-use-of-evidence/Overall-teacher-judgement-OTJ/A-student-s-achievement

Likelihoods of Scripts being Judged at Different Reporting Bands Depending on Text Level (Area Graph)



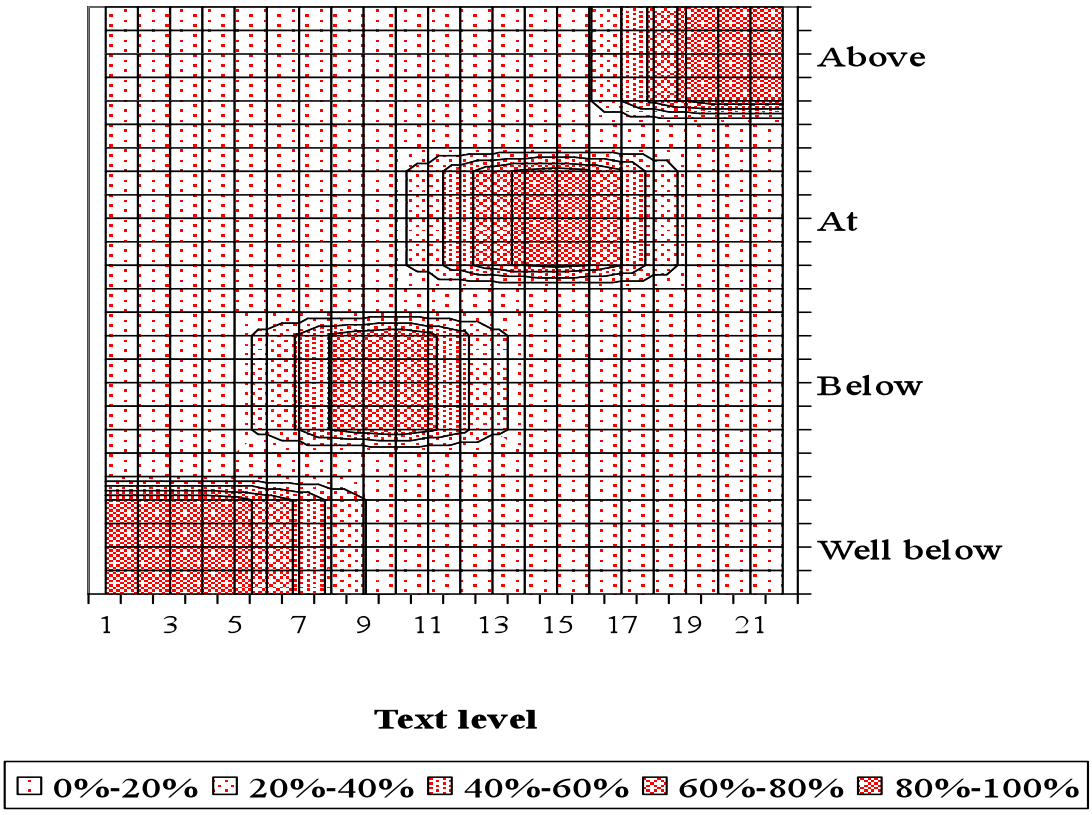
To read the graph, consider a vertical line above a text level of, for example, 11 (illustrated). Looking at the areas with which the line intersects, we can conclude the following.

- Almost no part of that line is in the blue area, reflecting that a script with a text level of 11 had almost no likelihood of being judged as ***above*** the 'after one year' national standard in reading.
- A medium sized part of that line is in the green area, reflecting that a script with a text level of 11 had a moderate likelihood of being judged as ***at*** the 'after one year' national standard in reading.
- A large part of that line is in the orange area, reflecting that a script with a text level of 11 had a large likelihood of being judged as ***below*** the 'after one year' national standard in reading.
- Only a very small part of that line is in the red area, reflecting that a script with a text level of 11 had a very small likelihood of being judged as ***well below*** the 'after one year' national standard in reading.

A teacher whose student had an Observation Survey text level of 11 could use the above information, together with their knowledge of other aspects of that student's reading performance to make an overall teacher judgement against the 'after one year' national standard in reading.

The graph below shows the same information, but now as a bird's eye view of the percentage of scripts in each of the four reporting bands against the text level. In it, brighter shades of red correspond to higher likelihoods of judgement.

Likelihoods of Scripts being Judged at Different Reporting Bands Depending on Text Level (Bird's Eye View)



The following points summarise the main features of the above graphs.

- 'Well below' judgements were somewhat likely[3] up to a text level of about 9. They were clearly the most likely[4] judgements up to a text level of about 7.
- 'Below' judgements were somewhat likely for text levels ranging from about 6 to 13. They were clearly the most likely judgements for text levels ranging from about 7 to 12.
- 'At' judgements were somewhat likely for text levels ranging from about 10 to 19. They were clearly the most likely judgements for text levels ranging from about 12 to 17.
- 'Above' judgements were somewhat likely for text levels of about 16 and higher. They were clearly the most likely judgements for text levels of about 18 and higher.

What about the stanines derived from the five different specific assessments included in the Observation Survey? One approach is to calculate an 'average stanine' and

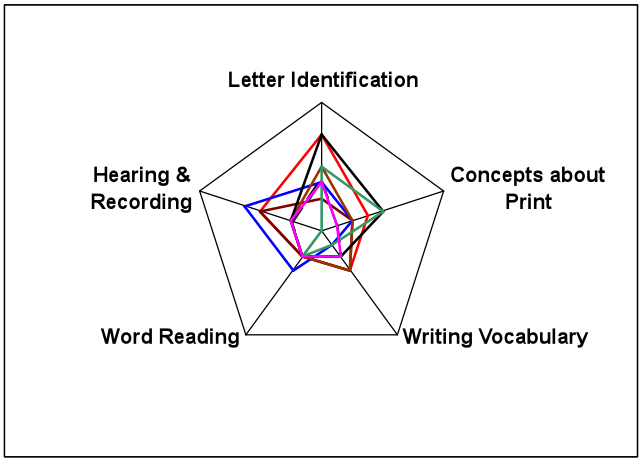
[3] i.e. with a probability of about 25% or more.
[4] i.e. with a probability greater than 50%.

relate this to the expert's judgements against the standards. If we do this, we come up with results something like this:

- 'Well below' judgements were somewhat likely up to an average stanine of about 5. They were clearly the most likely judgements up to an average stanine of about 4.4.
- 'Below' judgements were somewhat likely for average stanines ranging from about 3.6 to 6.4. They were clearly the most likely judgements for average stanines ranging from about 4.4 to 5.4.
- 'At' judgements were somewhat likely for average stanines ranging from about 4.4 to 7.4. They were clearly the most likely judgements for average stanines ranging from about 5.4 to 7.
- 'Above' judgements were somewhat likely for average stanines of about 6.6 and higher. They were clearly the most likely judgements for average stanines of about 7 and higher.

However, the five stanines relate to different aspects of reading achievement and an overall average may not be very informative when making a judgement against the standard. Another use of this kind of data is to present examples of actual results from scripts judged to be in each of the four reporting bands, as shown below.

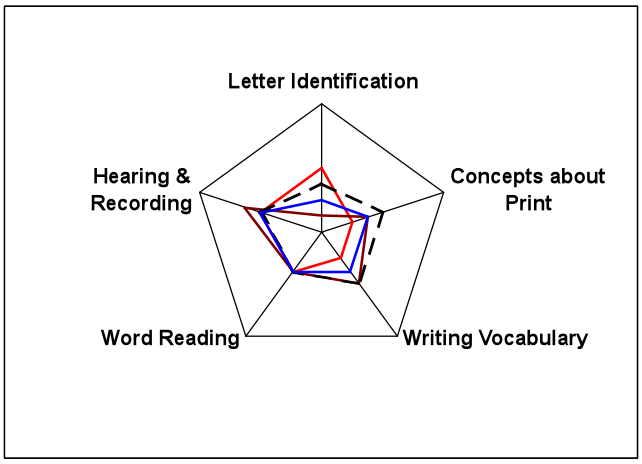
The 'star plots' below show, for scripts judged to be in each of the four reporting categories, the stanine values in each of the five assessments[5]. There are some interesting patterns to be observed, which are picked out alongside each plot.



Well Below

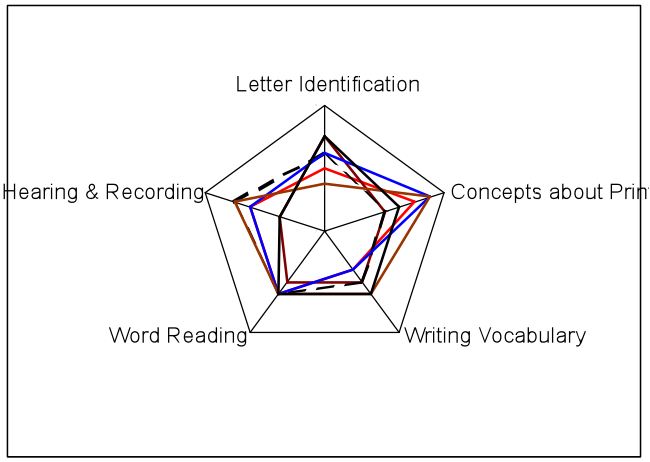
Variable results, but low on Word Reading and Writing Vocabulary

[5] Note that some assessments have maximum scores which correspond to more than one stanine, and in this case the lowest value has been taken in these plots. For Letter Identification and Hearing & recording the upper limit is taken to be stanine 7; for Word Reading it is 6.



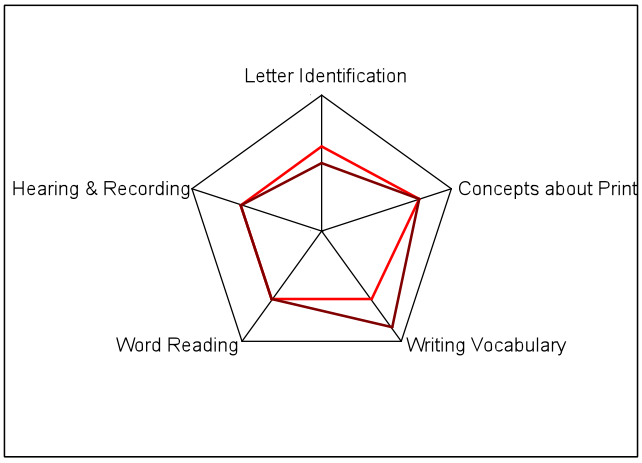
Below

More even performance, around stanine 5



At

Some variability, high stanines except Writing Vocabulary



Above

Particularly high on Writing Vocabulary

Teachers using the Observation Survey to make judgements against the 'after one year' reading standard may find this information useful.

Conclusion

This work will help teachers to use the Observation Survey as part of the evidence informing the overall teacher judgement of student performance against the 'after one year' national standard in reading. It is modest in scale and is yet to be replicated in a larger setting.

For each text level recorded on the Observation Survey, this work provides the likelihood of a student with that text level being judged as 'well below', 'below', 'at' and 'above' the 'after one year' national standard in reading. In addition, some indications of the likely relationships between stanine scores and teacher judgements are given. The fact that results are presented as likelihoods or probabilities reflects the fact that no one assessment tool will be enough to make a definite judgement against the reading standard. It highlights the need to use multiple pieces of evidence in informing overall teacher judgement of student performance against the national standards.

One thing you may notice when you study these results is that often there is not a consistent relationship between a test's norms (e.g. stanines, average scale scores for a year level etc.) and the most likely national standard reporting category. This is not a cause for concern. Test norms are based on what the average student of a given age *can* do; the standard relates to what all students *should be able* to do, if they are on track for a successful educational outcome. In some areas of learning, the two coincide – the average student (i.e. stanine 5) is at the required level. In other areas, there may be a general shortfall – only high-performing students (e.g. stanine 7+) are likely to reach the standard, with others needing to improve their achievement in order to do so. This is an important feature of national standards, and one of the ways in which they are intended to drive improved learning for all students.

Acknowledgements

We would like to thank all the people who have helped with and contributed to this work, especially the teachers and other literacy experts who took part in the script scrutiny event, and those who commented on initial drafts of this material to improve it, including members of the New Zealand Assessment Academy and AtoL directors. Thanks also to Professor Jeff Smith of Otago University, who inspired us to use the 'star plots' to present some of this data.

Important things to remember

Some key points to remember about this information:

- It is provisional, based on early analysis of existing data
- It is designed to help teachers make judgements against the standards on the basis of student performance on the Observation Survey
- It tries to capture the variability around the judgements made using any single tool, and shows the importance of pulling together different kinds of evidence to make an overall teacher judgement

- Results for more assessment tools are being processed and will be published as soon as possible
- Next year we will be publishing results based on more extensive data collection and analysis for a range of assessment tools
- Schools can use these results directly – they do not need to carry out their own 'script scrutiny' exercises.