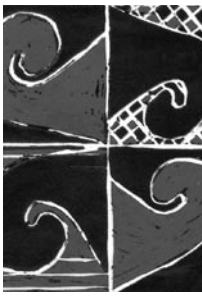
STAR Implications and Issues

An analysis of data generated from the *Supplementary Test of Achievement in Reading*



Dr Keith Greaney & Distinguished Professor William Tunmer
Massey University

ABSTRACT

In this paper we argue that the data generated from the New Zealand *Supplementary Test of Achievement in Reading* or *STAR* (Elley, 2001) may not be conducive to helping teachers identify specific learning needs at the individual level. We further contend that the main use of the data obtained from the *STAR* is the presentation of the results as aggregated stanines to show large group performance trends. To obtain a more in-depth understanding of the causes of low scores on the *STAR* we suggest that teachers might need to undertake an analysis of individual performances. Such an analysis was the focus of this paper and the results showed that for the two sub-tests analysed *(Sentence Comprehension* and *Vocabulary Knowledge),* the overriding problem was related to poor decoding ability. This was a surprising result given that all the students in the sample had scored 90% or better on the *Word Recognition* (STAR) sub-test.

Research
Keywords

Assessment tools, data collection, educational testing, needs assessment, reading achievement, reading comprehension, statistical analysis.

INTRODUCTION

> New Zealand teachers have access to more published procedures for the assessment of achievement in reading than in any other curriculum areas. The teaching of reading has been considered to be a strength of New Zealand primary schools and it is an area that has quite a strong tradition of teacher assessment (Education Review Office, 1999, p. 17).

It is often very difficult for teachers to identify specific teaching needs at the individual level on a reading test that is undertaken in a silent reading situation. The main reason for this is that the metacognitive and cognitive strategies that the student uses to complete such a test are not transparent in the test responses (i.e., answers). This is more problematic with tests such as the *Supplementary Test of Achievement in Reading* or *STAR* (Elley, 2001) where the answers are presented in multiple choice formats in which the student is only required to select one answer (word or letter) from a selection of choices. The quality of the data that teachers have to work from as a basis for *diagnosing* specific learning problems and/or planning for future teaching is therefore compromised.

Even the presentation of stanine scores (which are obtained from converting the raw score totals) do not offer particularly useful diagnostic information at the individual level of performance. Yet, the stanine scores are the most often cited *STAR* data.

SOME RELEVANT RESEARCH STUDIES

> If we are really going to help students, we need to understand the underlying reasons for their test failure. Simply knowing which students have failed is a bit like knowing that you have a fever when you are feeling ill but having no idea of the cause or cure. A test score, like a fever, is a symptom that demands more specific analysis of the problem (Valencia & Buly, 2004, p. 520).

In a study that investigated why 4th grade (New Zealand Year 5) students had failed the mandated *Washington Assessment of Student Learning (WASL)* reading test, Buly and Valencia (2002) drew attention to the problem of using group-administered tests to identify reading difficulties at the individual level. These authors noted for example that 'from the scores alone, derived from a group administered measure, little was known about the specific reading capabilities or difficulties that may have contributed to their poor performance' (p. 124).

In order to better understand why some students performed poorly on a mandated reading test, Valencia and Buly (2004) undertook an analysis of test data and identified six distinct sub-groups or profiles of reading disability. These six sub-groups were labelled as;

- automatic word callers (18%)
- struggling word callers (15%)
- word stumblers (17%)
- slow comprehenders (24%)
- slow word callers (17%)
- disabled readers (9%).

It is interesting to note that almost 50% of the error sub-groups were directly attributable to poor or inefficient word identification skills. Many of the remaining sub-groups were also most likely to have word identification problems. The authors noted for example that the disabled readers had 'experienced severe difficulty in all three areas: word identification, meaning and fluency' (Valencia & Buly, 2004, p. 527).

However, while many of the students had experienced word identification problems, this does not mean that every poor performing student will necessarily require more instruction in word identification strategies. In-depth analyses of test data at the individual level are more likely to show that "one size of instruction" does not necessarily fit all students. Students fail tests for a variety of reasons and it is important that teachers are aware of these different reasons when attempting to plan and teach programmes.

Dewitz and Dewitz (2003) investigated the reasons why some 4th and 5th grade (year 5-6 New Zealand) students had failed the mandated Qualitative Reading Inventory *(QR1-3).* Some of the reasons why the students had scored poorly on this test included a failure to adequately link key ideas in the passages and to overly rely on prior knowledge when answering the questions. In their investigation the authors had asked the students to discuss and elaborate their particular responses to the various test questions. These elaboration opportunities allowed the students to explain their reasoning in much greater detail than was evident by only checking their answers at a later date. The investigators were also more able to identify inefficient or inappropriate metacognitive strategies that the students had used and which had subsequently resulted in or contributed to their receiving a low test grade score. A similar study by Greaney (2004) involved a retrospective analysis of some New Zealand *Progressive Achievement Test* (reading comprehension) data to identify the reasons why some students scored poorly on this test. Twenty-nine Year 4-6 students were asked to re-read some of the passages from their reading comprehension test and to explain the reasons why they had selected the incorrect (multiple choice) answers. Like the Dewitz and Dewitz (2003) study, Greaney (2004) also noted that the students often used inappropriate prior knowledge and showed an inability to link key ideas within the passages. However, the important point to be taken from these studies is that teachers need to be aware that raw test score data by themselves is of only limited value, particularly if used as a basis for identifying the teaching needs of students who have low scores.

THE NEW ZEALAND *SUPPLEMENTARY TEST OF ACHIEVEMENT IN READING (STAR)*

Although there are no national data on the use of this test, anecdotal records suggest that it is used in more than 50% of New Zealand primary schools at least once per year. The *STAR (Year 4-6)* test has national norms for three different test-points during the year (February-May, June-August, September-November) and comprises four sub-tests; *Word Recognition, Sentence Comprehension, Paragraph Comprehension* and *Vocabulary Knowledge.* However, the *Paragraph Comprehension* sub-test was not a focus for this paper, so will not be discussed further.

Word Recognition

According to the *Teachers' Manual* (Elley 2001) the *Word Recognition* sub-test 'shows how well the pupil can decode words that are familiar in their spoken language' (p. 7). Ten groups of four words are presented and the student is required to first, identify a picture (e.g., an *umbrella)* and then identify the matching target word from the group (e.g., *under, union, umbrella, umpire).* The maximum score for this sub-test is 10.

Sentence comprehension

The *Teachers' Manual* states that the *Sentence Comprehension* sub-test 'shows how well pupils can read for meaning' (p. 7). Ten sentences are presented and the student is required to select one word (from groups of four) to complete each sentence (e.g., *Your mother's sister is your…* **uncle / aunt / grandmother / cousin**). In all cases, the missing word is at the end of the sentence. The maximum score for this sub-test is 10.

Vocabulary Knowledge

According to the *Teachers' Manual* the *Vocabulary Knowledge* sub-test measures 'pupils' knowledge of word meanings in context' (p. 7). Simple sentences are presented with target words underlined and the pupil is required to select (from a group of four alternatives), the word that is closest in meaning to the underlined word (e.g., *The lake is often* calm *before a storm* – **smooth / rough / dark / beautiful**). The maximum score for this sub-test is 10.

How the STAR test scores are reported

Each of the sub-test scores are added together (maximum 50, as the *Paragraph Comprehension* subtest is marked out of 20) and the total raw scores are then converted to stanine scores. Many schools prefer to use the stanine score as the standard measure for reporting individual and class level performance data.

METHOD

The Research Problem

The most common measure used by most schools for reporting *STAR* reading test achievement data is the *stanine* unit. We contend that this unit is not refined enough for teachers to gain an in-depth understanding of *causes* of low performances on this test. While a very high stanine score (e.g., 8 or 9) may suggest proficiency of the skills assessed, a low stanine score of 4 or less may not necessarily indicate anything except that the student has scored low. Even looking at individual sub-test raw score data may not necessarily shed light on the *reasons* why some students score poorly. Yet schools seldom report or analyse the *STAR* data beyond the level of the stanine unit.

Procedures

Twenty-one Year 4-6 students who had scored 90%+ (9 or 10) on the *Word Recognition* STAR sub-test but 50% or less on the *Sentence Comprehension* and/or the *Vocabulary Knowledge* sub-tests were selected to participate in the study. The students were required to re-read orally the items in the sub-tests that they had incorrectly selected in their original test. Records of their oral reading of the test items were recorded to check the level at which the students were able to accurately identify (decode) the words in the various sub-test tasks.

Rationale for Selecting Students who Scored 90% or Better on the Word Recognition Sub-test

According to the *Teachers' Manual* pupils who score 90% or more on the *Word Recognition* sub-test have 'mastered the elements of decoding' (p.19). This statement therefore suggests that provided a student scores 90% or more on this sub-test, any low scores on the remaining sub-tests are not likely to be seen to be caused by a decoding problem. Teachers would also therefore, be unlikely to even consider investigating decoding problems with this group of students.

ANALYSIS

The 21 Year 4-6 students' raw scores for the three STAR focus sub-tests (and the stanines) are presented in Table 1.

Table 1

Summary of Individual STAR Sub-Test Scores

Id No	Word Recognition (n=10)	Sentence Comprehension (n=10)	Vocabulary Knowledge (n=10)	Stanine
1	10	9	(4)	7
2	10	10	(4)	7
3	9	7	(5)	6
4	9	7	(5)	6
5	9	7	(5)	6
6	9	6	(4)	4
7	10	8	(3)	5
8	9	9	(5)	5
9	9	6	(5)	5
10	9	6	(4)	6
11	9	8	(4)	6
12	9	(5)	(2)	5
13	9	(4)	(2)	3
14	9	(5)	(2)	4
15	9	(4)	(5)	4
16	9	(5)	(5)	4
17	9	(5)	(3)	4
18	9	(4)	(5)	4
19	9	(5)	(2)	4
20	9	(5)	(5)	4
21	10	(5)	6	6

The results in this table show several interesting points. First, all 21 students have scored 90% or better on the *Word Recognition* sub-test which, (according to the *Teachers' Manual)* would all be categorised as "efficient decoders". A score of 90% or better on this sub-test was the initial "trigger" for selection for the study. The target scores that were analysed are inserted within brackets. That is the scores of 50% or less on either or both of the *Sentence Comprehension* and *Vocabulary Knowledge* sub-tests. Second, the data in the table show that all but one of the students (i.e., 95%) had scored 50% or less on the *Vocabulary Knowledge* sub-test, and ten students (i.e., 47%) scored below the threshold score for the *Sentence Comprehension* sub-test. Third, nine students (i.e., 42%) scored below the threshold levels on **both** the *Sentence Comprehension* and *Vocabulary Knowledge* sub-tests. The results of the analysis of the responses for the *Sentence Comprehension,* and *Vocabulary Knowledge STAR* sub-tests are presented in Tables 2-3. *The Paragraph Comprehension* sub-test scores have not been included in Table 1 as this sub-test was not the focus of the study. However, each student's stanine score is included and it is interesting to note that all but one student have stanine scores between 4 and 7.

Given that schools tend to use the stanine score as the main measure for reporting the STAR results, and given that the STAR Manual (see p. 17) reports that stanines 4 to 7 fall into the "average" category, these students are not likely receive specific attention on the basis of these scores.

Analysis of the *Sentence Comprehension* Sub-Test Data

The data in Table 2 present a summary of the two main categories of errors that affected the students' performances on the *Sentence Comprehension* sub-test. The data also show that two sources of difficulty were apparent for the students. These included poor decoding skills and poor vocabulary knowledge. Because the *Sentence Comprehension* sub-test required the students to select a word to complete each sentence, it was important that the words (and especially the target words) were correctly decoded first. If all the words in the sentences (including the four optional words) were read correctly yet an incorrect answer was selected, this was coded as a vocabulary error. However, if the student was unable to correctly identify a *key* word in the sentence or was unable to read the target word in the answer options, then this was coded as a decoding error.

Table 2

Main Error Categories for the Sentence Comprehension STAR Sub-Test

Teaching focus shown by asterisk (*)

Id No	Word	Sentence	Main Error Category	
	Recognition Score	Comprehension Score	Decoding	Vocabulary
12	9	(5)	5*	0
13	9	(4)	6*	0
14	9	(5)	4*	1
15	9	(4)	4*	2
16	9	(5)	3*	2
17	9	(5)	3*	2
18	9	(4)	5*	1
19	9	(5)	5*	0
20	9	(5)	4*	1
21	10	(5)	2	3*

The analysis of the students' main reasons for their errors on the *Sentence Comprehension* sub-test indicate that, in all but one case, they had problems decoding a key word either in the initial sentence or the target word among the options in the answers. For example, in Table 2 student (Id 12) scored 5 (out of a possible 10) for the *Sentence Comprehension* sub-test. On closer analysis of the errors it was found that this student was unable to read either the key focus word from among the four answers or one of the key content words in the question. As an example, in the sentence: *As the snake slithered closer, Sam's eyes grew wide with …* **fascinating / pleasure / anticipation / terror,** several students were unable to correctly decode the word slithered and others could not identify the target word terror. While some of these students may well also lack the vocabulary knowledge (which is the main skill that this sub-test assesses), the data in Table 2 suggest that their *main* problem is poor decoding skills. This was an interesting (and unexpected) finding given that these students had all scored 90% or better on the *Word Recognition* sub-test. On the basis of this data, the main teaching focus for most of these students should be one that includes an emphasis on developing more efficient decoding skills. This is because 73% (i.e., 41) of the errors were related to poor decoding skills.

Analysis of the *Vocabulary Knowledge* Sub-Test Data

The results for the analysis of the *Vocabulary Knowledge* sub-test are presented in Table 3. As with the *Sentence Comprehension* sub-test, the purpose of the analysis of the *Vocabulary Knowledge* sub-test was to investigate the extent to which errors reflected vocabulary knowledge problems or decoding problems.

If the student was able to correctly identify all the words both in the sentences and (in particular) the target word among the four alternates yet selected an incorrect answer, the cause for the error was categorised as a vocabulary problem. A decoding error was recorded whenever the student was unable to correctly decode a key word in the sentence or the key target word among the four alternates. For example, in the sentence; *I found Nick's story quite incredible.* **convincing / inaccurate / memorable / unbelievable,** the key words included; found, story, incredible and unbelievable. It was hypothesised that a decoding error involving any of these words would be likely to impact negatively on the student's ability to select the correct answer option.

The data in Table 3 show that 20 students scored 50% or less in the *Vocabulary Knowledge* sub-test. There were a total of 121 errors analysed from the *Vocabulary Knowledge* sub-test and 60% (i.e., 73) of these errors were related to poor decoding skills with the remaining 40% categorised as vocabulary-related. In this particular sub-test analysis it would be expected that the teacher plan for two distinct teaching foci, where one group of 11 students would require a decoding skills focus and the other nine students would benefit from an instructional approach that encourages the development of vocabulary skills.

Table 3

Main Error Categories for the Vocabulary Knowledge Sub-Test

Teaching focus shown by asterisk (*)

Id No	Word Recognition	Vocabulary Knowledge	Main Error Category	
	Score	Score	Decoding	Vocabulary
1	10	(4)	0	6*
2	10	(4)	2	4*
3	9	(5)	4*	1
4	9	(5)	5*	0
5	9	(5)	0	5*
6	9	(4)	6*	0
7	10	(3)	5*	2
8	9	(5)	4*	1
9	9	(5)	4*	1
10	9	(4)	2	4*
11	9	(11)	2	4*
12	9	(2)	8*	0
13	9	(2)	7*	1
14	9	(2)	3	5*
15	9	(5)	1	4*
16	9	(5)	3*	2
17	9	(3)	3	4*
18	9	(5)	4*	1
19	9	(2)	8*	0
20	10	(5)	2	3*

GENERAL DISCUSSION

The *STAR Teachers' Manual* (Elley, 2001) states that 'most students with low scores on *Word Recognition* have low scores on all sub-tests' (p.19). This would be expected because if students were unable to read the words adequately, they would certainly also have problems completing the *STAR* reading test. However, the focus of the current study included the students who were *very good decoders* (as evidenced by their **high** *Word Recognition* sub-test scores). Twenty one "high decoder" students had scored 50% or less on the *Sentence Comprehension* and/or the *Vocabulary Knowledge* sub-test. These students were given the opportunity to orally re-read their incorrect test items on the *STAR* test in order to investigate the level at which poor decoding skills may or may not have impacted on the low scores. The data showed that the main cause of their low scores for both the *Sentence Comprehension* and *Vocabulary Knowledge* sub-tests were due to poor/inefficient decoding skills. This was rather surprising given that the students were initially selected based on their very high *Word Recognition* sub-test scores. Teachers may therefore be led to believe that a low *Vocabulary Knowledge* score is due to only poor vocabulary knowledge. While some of the students may well have also had low vocabulary knowledge, the investigation demonstrated that many of the students were simply unable to decode the target words adequately.

In these cases, a teaching focus that included explicit instruction in how to decode such words would be more relevant than one that focused on developing vocabulary knowledge. Yet most teachers would be more likely to introduce a vocabulary, rather than a decoding focus for students who score poorly on the *Vocabulary Knowledge* sub-test. Similarly, while the results show that the main problem with the Sentence Comprehension sub-test was also due to poor decoding, it is unlikely that teachers would focus on this problem when deciding on a teaching focus. It is also interesting to note that the *STAR Teachers' Manual* does not highlight the possibility that poor decoding skills may still be a likely problem on sub-tests other than Word Recognition.

CONCLUSIONS

Timperley and Parr (2004) discuss the benefits of analysing assessment data in some depth in order that teachers develop a more focused understanding of the specific teaching needs. The authors term this analysis process "mining the data". However, in the text the authors use the *STAR* reading test data as an example of mining the data but they also focus mainly on the use of stanines (pp. 88-89) to show large group results. They do not discuss the relevance for "mining the data" in a way that shows *individual* performance problems.

The *STAR* reading test is used in many schools in New Zealand, and often in preference to the *Progressive Achievement Tests (PATs).* Moreover, schools generally prefer to report the results by presenting stanine scores for large groups. While this level of data presentation is suitable for showing large group trends, we maintain that the stanine statistic is not refined enough to reveal the underlying cause of low scores. To obtain information about causation we recommend that teachers undertake a finer-grained analysis of student performances at the individual level.

Many normed reading tests (e.g., *STAR, PAT*) include multiple choice answer responses. However, such tests are not particularly informative as *diagnostic* measures. There are many reasons why a student selects incorrect responses in multiple choice tests and it is important that teachers are aware of these possibilities if they are to design focused interventions based on individual needs.

Finally, it is important to note that the *STAR* reading test is only one reading assessment tool, and when taken in isolation, may not be particularly informative. However, in this study we demonstrated that, when further more in-depth analyses were undertaken with students whose scores were particularly low, that useful and more relevant information was obtained about the *causes* of the low scores. Furthermore, this additional information was not transparent merely from viewing the aggregation of stanine scores alone. As with any standardised reading assessment tool, it is useful for teachers to be aware of the limitations of using grouped or aggregated data as a proxy measure for causation. Popham (2003) also warns that most accountability tests have very little value for improving teaching and learning. In fact Popham argues that such tests can even 'lull educators into believing that they have appropriate data when they do not, and as a consequence, many educators fail to ask for more meaningful, instructionally valuable data that would help them teach students better' (p.192).

It is important that if teachers are sincere about "closing the reading achievement gap" between the high and low ability readers in their class, they need to be prepared to undertake a finer-grained analysis of the assessment data *for their low-achieving students* in order to develop more focused teaching strategies for this group. Such an analysis was the focus in this paper.

REFERENCES:

Buly, M., & Valencia, S. (2002). Below the Bar: Profiles of students who fail state reading assessments. *Educational Evaluation and Policy Analysis, 24*(3), 219- 239.

Dewitz, P., & Dewitz, P. (2003). They can read the words but they can't understand: Refining comprehension assessment. *The Reading Teacher, 56*(5), 422-435.

Elley, W. B. (2001). *Supplementary Test of Achievement in Reading.* Wellington, New Zealand: New Zealand Council for Educational Research.

Education Review Office (1999). *Student assessment: Practices in primary schools,* 2, Winter, Crown copyright.

Greaney, K. T. (2004). Factors affecting reading comprehension performance: A retrospective analysis of some year 4-6 data. *New Zealand Journal of Educational Studies, 39*(1), 3-22.

Popham, W. J. (2003). The seductive allure of data. *Educational Psychology, Annual Edition* (pp. 192-195). Dubuque, IA. McGraw-Hill.

Timperley, H., & Parr, J. (2004). *Using evidence in teaching practice: Implications for professional learning.* Auckland, New Zealand: Hodder Moa Beckett.

Valencia, S., & Buly, M. R. (2004). Behind test scores: What struggling readers really need. *The Reading Teacher, 57*(6), 520-531.

Keith Greaney

AUTHOR PROFILE

Dr Keith Greaney is a Senior Lecturer in the Department of Learning and Teaching at Massey University College of Education. Before coming to Massey, Keith was a primary school teacher for 28 years, including two years as a special class teacher and 12 years as a Resource Teacher: Reading. He is also a trained Reading Recovery teacher. Keith teaches a paper in the post-graduate Diploma in Literacy Education course and assists with the supervision of students undertaking Masterate research in literacy-related areas.

Email

K.T.Greaney@massey.ac.nz

Bill Tunmer

AUTHOR PROFILE

Professor Bill Tunmer is a Distinguished Professor of Educational Psychology at Massey University College of Education where he teaches and researches in the areas of reading disabilities.