



Analysis of The Early Literacy Observation Survey

Patrick Griffin
Masa Pavlovic
Esther Care

Assessment Research Centre
University of Melbourne
July 2007

The Faculty of Education
The University of Melbourne Victoria 3010 Australia
Telephone +61 3 8344 8206 Fax + 61 3 8344 8790
<http://www.edfac.unimelb.edu.au/ARC>



Index of Contents

The Early Years Literacy Preliminary Norming Project	3
Concepts about Print	3
Letter Identification	3
Word Reading	4
The Burt Word Reading Test	4
Writing Vocabulary	4
Hearing and Recording Sounds in Words.....	5
The Preliminary Norming Trials.....	5
Test Distributions for All Students	6
Test Distributions by Grade Levels	9
Item Response Analyses	13
Summary	21
Conclusions and Recommendations	22
Summary of Recommendations	23
Sample Sizes	26
Reference	26

The Early Years Literacy Preliminary Norming Project

This project examines the series of tests used with Prep to Grade 2 students to examine what they understand about print, and their beginning reading strategies. It represents a preliminary attempt to establish the adequacy of the data available for norming the early years literacy assessment protocols.

Six tests are used:

1. Concepts About Print
2. Letter Identification
3. Word Reading
4. The Burt Word Reading Test
5. Writing Vocabulary
6. Hearing and Recording Sounds in Words.

Concepts about Print

In *Concepts about Print* the teacher chooses one of four booklets to use with the student. The four booklets are titled *Sand, Stones, Follow Me Moon* and *No Shoes*. In this test the teacher attempts to identify what the student knows about printed language. The teacher explores with the student ideas of directionality, letters, words, spaces, punctuation and its meaning, order of words, letter case and directionality of left to right, parts of the book and the way in which a story is organised. The student is expected to know that it is the print, not the picture that conveys the message; to know where to start, and which way to go with the print; to be able to match words to sounds and identify the difference between first and last in concepts. They need to be able to find the beginning and end of stories and lines and be able to see that the line order is important; to know that the left page is read before the right page, that the question mark has a meaning, and that the full stop has a meaning as does a comma and quotation marks. They need to be able to locate and identify specific letters in the stories and they need to be able to identify specific words within the story. The scoring rules for the test are laid out in Marie-Clay's book "*An observation survey of early literacy achievement*". By and large the scoring rules are clear and concise and in this particular test, there is very little judgement required. The scoring rules are objective.

Letter Identification

The second test is *Letter Identification*. In this the student is presented with the alphabet in upper and lower case, as well as the letters 'a' and 'g' in two commonly encountered formats giving a total of 54 symbols the student is expected to recognise. Each letter is presented to the student in order, first as capitals and then in lower case. The task consists of the student naming the letter and/or giving a sound for the letter and being able to state a word that begins or starts with or sounds like that letter. Scoring is totalled using all three responses added together, so that it could be interpreted that there is a total of 3 points for each symbol; however the whole item is

scored a 1 or 0. This could be confusing but, in particular, the rubrics could be improved to extend the test and the way it is interpreted. The total score for a perfect overall response is given in the guide booklet as a score out of 54, that is, a student gets a score of 1 if they are able to name the letter or give a sound or give a word that starts with it. It would be interesting to trial the scoring with naming, sounding and using the letter to see whether or not it produces a different result to those reported in these analyses. However, this is a letter recognition test and not a letter use test, so the validity would have to be considered.

Word Reading

The third test is a word recognition task. To administer this test the teacher is expected to choose one of three alternative lists, each list contains 15 words, randomly selected from the 45 most commonly used words in the NZ series of *Ready To Read* booklets. The lists are considered to be equivalent in terms of difficulty and it is up to the teacher to select the list of words they intend to use. The equivalence of the lists of words perhaps needs to be evaluated.

The Burt Word Reading Test

The fourth test is the NZ revision of the *Burt Word Reading Test*. This is a test of 110 words increasing in difficulty. The teacher records which words the student recognises and the student gets a score out of a total of 110.

Writing Vocabulary

The fifth test is Writing Vocabulary. The student is asked to write down all the words they know. They are given 10 minutes to do so. The teacher records the word as written if the spelling is correct; if the child knows what the word is meant to be; if the writing is from left to right; if it is clear what the letters are meant to be. A series of words is scored as multiple words and a mixture of capitals and lower case is not detrimental to the score. The teacher is allowed to prompt and specific prompts are given in the manual. Students are first of all asked to write their name; if they are not able to do that they are prompted to write the word 'is', 'to' or the word 'I'. They could be prompted with the words 'go', 'me', 'look' or 'come'. Other prompts might be questions that the teacher asks the student, such as "Do you know any other children's names", "Can you write things that you do", "Can you write things that you have in your house", "Can you write things that you ride on", "Can you write things that you eat". One point is given for each word that meets the criteria.

Hearing and Recording Sounds in Words

The last of the tests is *Hearing and Recording Sounds in Words*. Five forms of the test are given. Each consists of a sentence or two sentences containing a total of 37 phonemes. Most of the phonemes are letters of the alphabet, but they can include the phonemes ‘th’, ‘sh’ or ‘oo’. The forms of the test consist of sentences.

Form A: “I have a big dog at home. Today I am going to take him to school”.

Form B: “Mum has gone up to the shop. She will get milk and bread”.

Form C: “I can see the red boat that we are going to have a ride in”.

Form D: “The bus is coming. It will stop here to let me get on”

Form E: “The boy is riding his bike. You can go very fast on it”.

Each of these forms contains 37 phonemes. The teacher simply records each word that is written correctly.

The Preliminary Norming Trials

Data from 137 children were obtained. Of these 61 were boys and 49 were girls. There were 10 five year olds, 23 six years olds, 15 seven year olds and 4 eight year olds. A total of 52 of the 137 were able to provide age data. There were 44 children in Prep, 54 in Grade 1 and 39 in Grade 2. These sample sizes were clearly too small for establishing norms.

Concepts about Print was used with 133 students; *Letter Identification* with 137; *Word Reading* with 137; *The Burt Word Reading Test* with 132; *Writing Vocabulary* with 137, and *Hearing and Recording Sounds* with 136. These data are reported in Table 1.

Table 1: Descriptive frequencies for each test

	N	Mean	Std. Deviation	Skew	Std. Error Skew	Min	Max
Concepts Print Total	133	17.14	5.361	-.842	.210	3	24
Letter Identification Total	137	45.15	15.114	-1.987	.207	0	54
Word Reading Total	137	8.72	5.914	-.282	.207	0	15
Burt Reading Total	132	23.05	20.426	.856	.211	0	88
Writing Vocabulary Total	137	26.55	21.368	.674	.207	0	99
Hearing & Recording Total	136	26.29	12.988	-.973	.208	0	37

Table 2 shows the number of students in Prep, Grade 1 and Grade 2 who obtained the maximum possible scores (as identified by bolded scores in Table 1) for those tests for which total scores are relevant. That students can obtain the maximum total scores indicates a test ceiling effect.

Table 2: Frequencies of maximum scores obtained by Grade levels across tests

	Prep (N = 44)	Grade 1 (N = 54)	Grade 2 (N = 39)
Concepts Print Total	0	7	8
Letter Identification Total	3	20	18
Word Reading Total	2	15	19
Hearing & Recording Total	2	18	20

Test Distributions for All Students

Figure 1 illustrates the score distribution for the *Concepts about Print* test. An overall mean of 17 with a standard deviation of 5.4 was identified for 133 students.

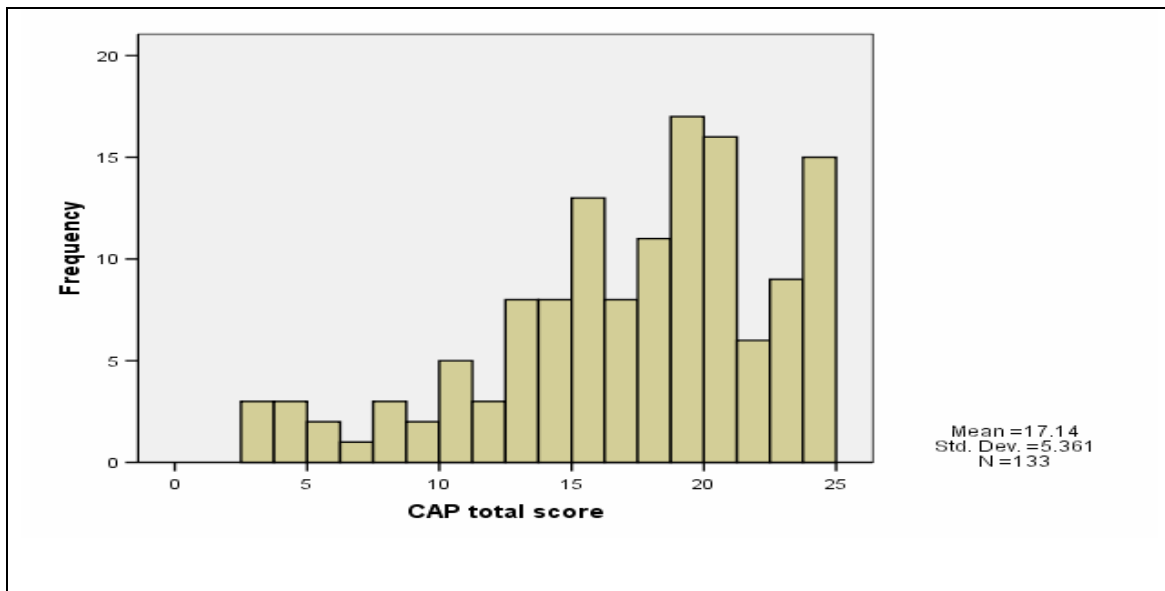


Figure 1: Score distribution for *Concepts about Print*.

Figure 2 shows the total scores for the *Letter Identification* test, clearly illustrating that a ceiling effect was in force. An overall mean of 45 with a standard deviation of 15 was identified for 137 students.

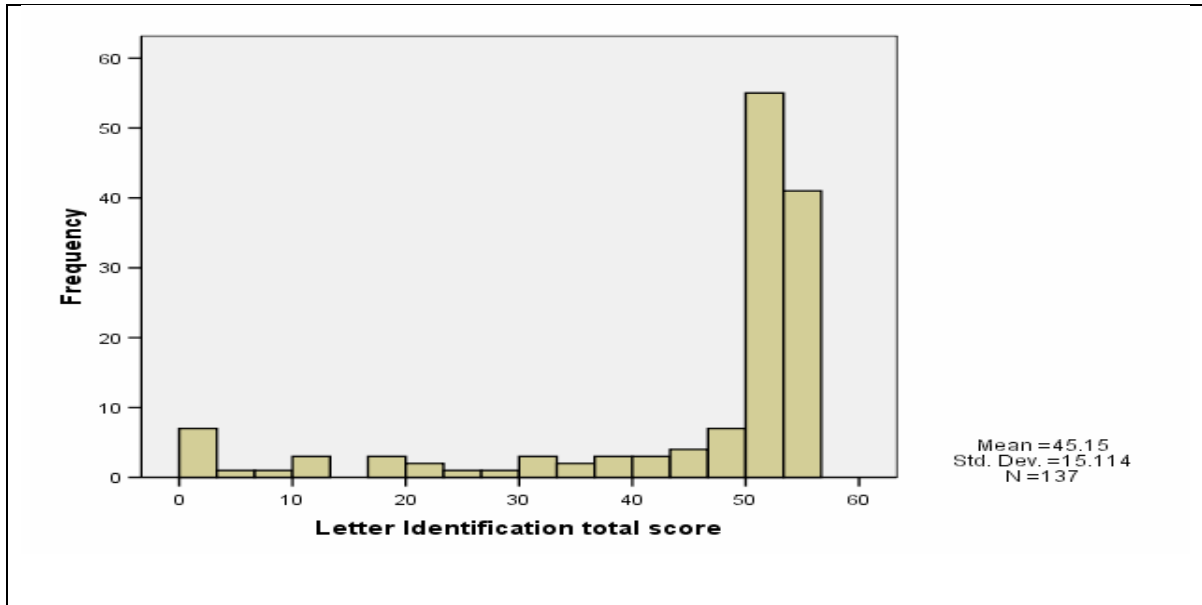


Figure 2: Score distribution for *Letter Identification*

Figure 3 illustrates the *Word Reading* test score distribution. Again, out of a total possible maximum score of 15, the average was 8.7 with a standard deviation of almost 6. It is a bimodal distribution which tends to illustrate the rapid progress occurring with this skill.

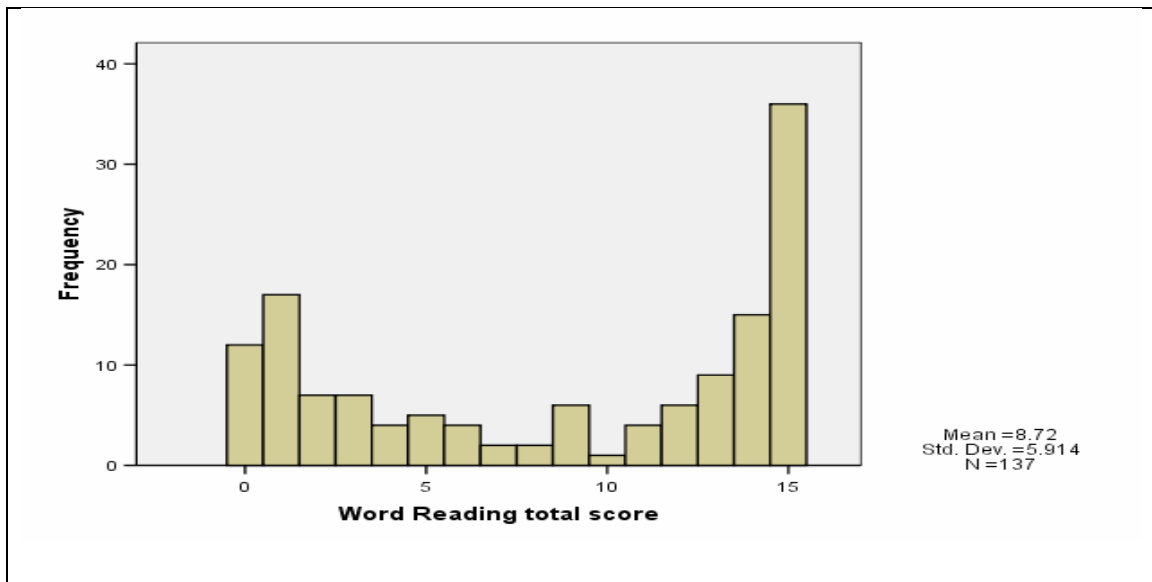


Figure 3: Score distribution for *Word Reading*

Figure 4 presents the *Burt Word Reading Test* total score distribution. This distribution is skewed towards the lower end with a mean of 23 and standard deviation of 20 across the 132 students. There is clear opportunity for growth to be measured with this test.

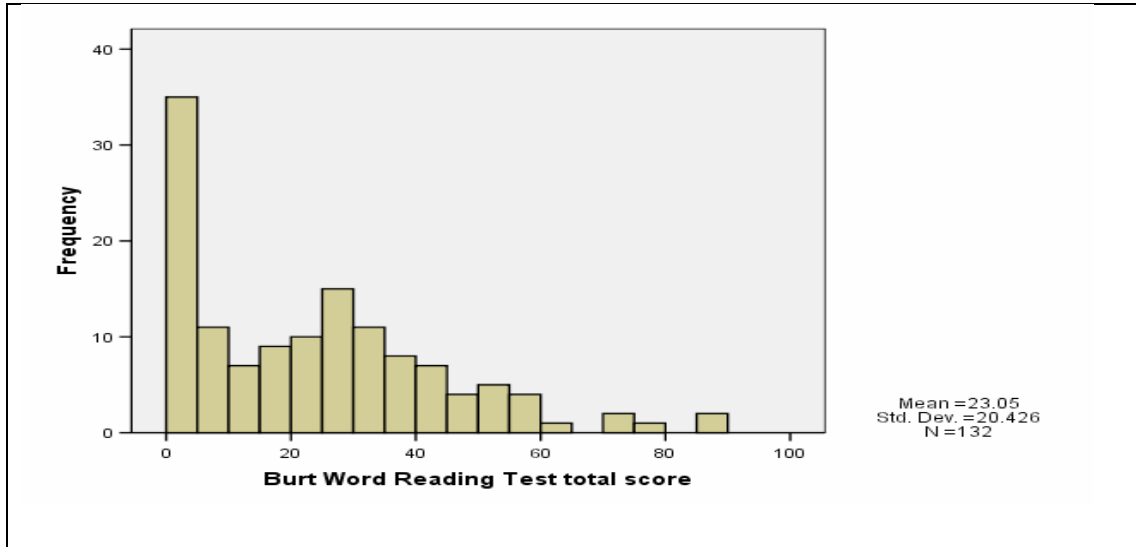


Figure 4: Score distribution for the *Burt Word Reading Test*

Figure 5 presents the *Writing Vocabulary* distribution of scores. This illustrates room for monitoring student improvement or growth.

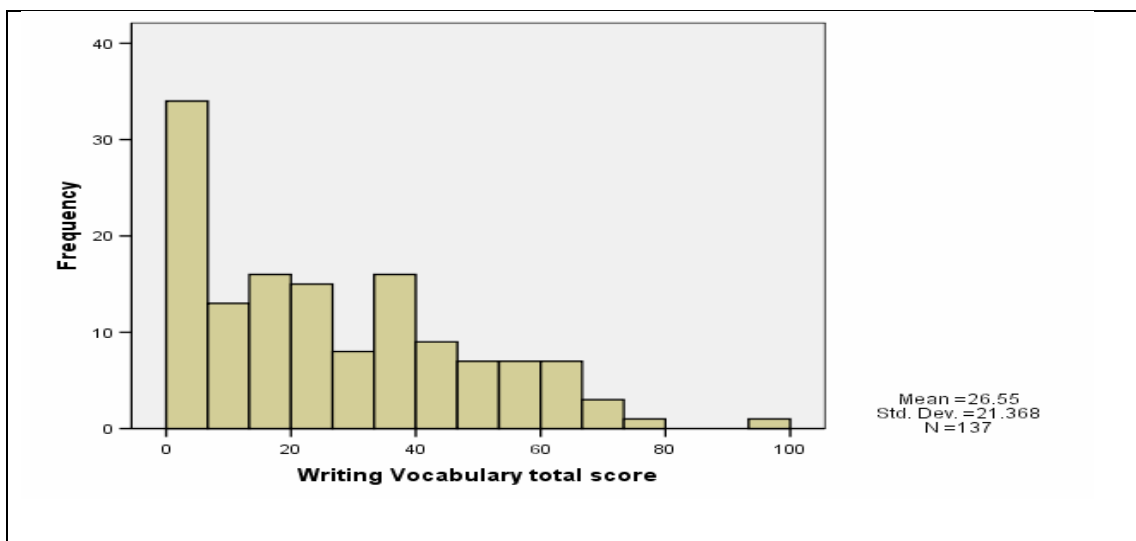


Figure 5: Score distribution for *Writing Vocabulary*

Figure 6 presents the data on *Hearing and Recording Sounds in Words*. In this Figure it can be seen that there is a relatively uniform distribution except that there is a large number of students who score very high on the test. Many score at the test ceiling (37) leading to a mean of 26 and a standard deviation of 12.

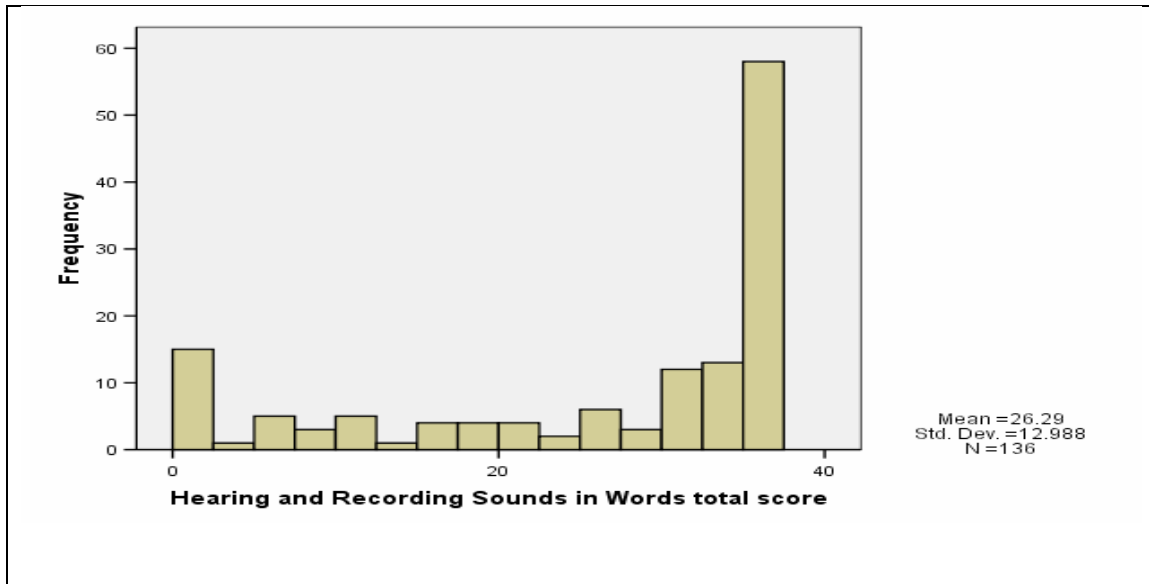


Figure 6: Score distribution for *Hearing and Recording Sounds in Words*

Test Distributions by Grade Levels

When the six tests are broken down by Grade level, various patterns begin to emerge. For the Prep students we can see that the *Concepts about Print* has a fairly uniform distribution. *Letter Identification* is skewed more towards the upper end, showing a median of 48 but with a long tail in the distribution indicating that some of the Prep students at this time of the year, or time of testing, have very few of the skills being sampled. The *Word Reading* total score at Prep supports the notion that the reading skills are low at that level as indicated similarly by the *Burt Reading Test*, the *Writing Vocabulary Test* and the *Hearing and Recording Sounds in Words* test.

In Figure 7 the distributions are presented for the six tests using box and whisker plots for the Prep children. It can be seen from this Figure that *Concepts about Print*, the *Burt Word Reading Test*, the *Writing Vocabulary* test, and *Hearing and Recording Sounds* test all have some room to show growth and development.

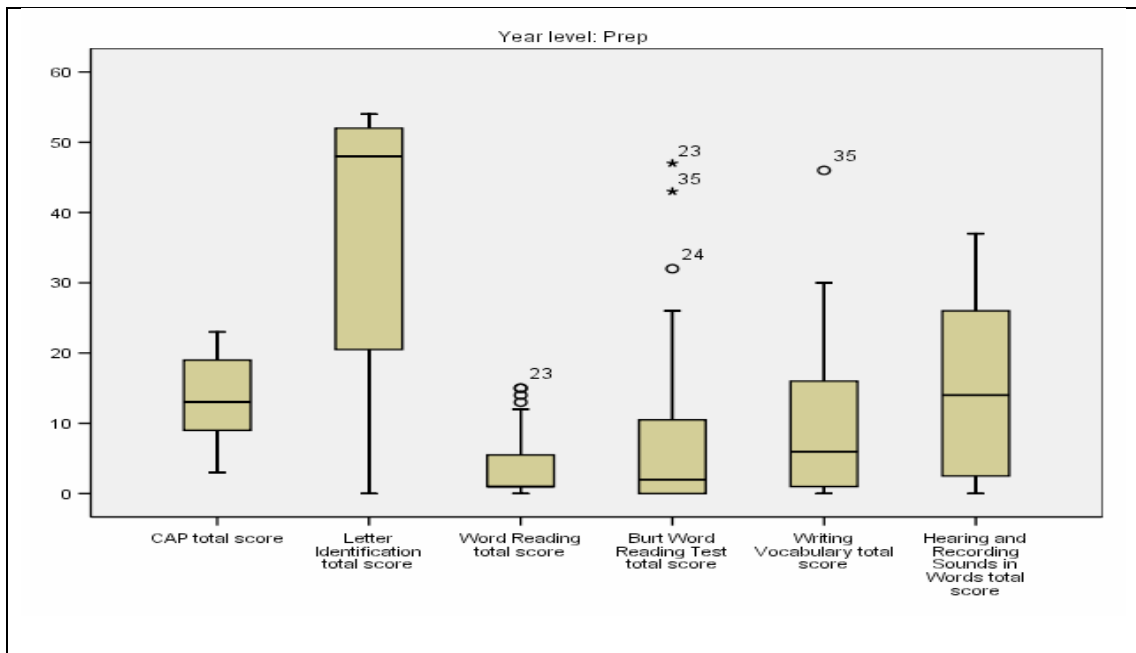


Figure 7: Test distributions for Prep level

Figure 8 shows parallel information for Grade 1. It can be seen that there is a rapid shift in the distribution of scores for the *Concepts about Print* in that most students have moved towards the upper end of the distribution and in the *Letter Identification* test most students have scored the maximum. Very little change has occurred in the *Word Reading* total score. The *Burt Word Reading Test* has now a much wider distribution as has *Writing Vocabulary*, but *Hearing and Recording Sounds* (phonemic tests) have moved rapidly towards the upper ceiling of the test.

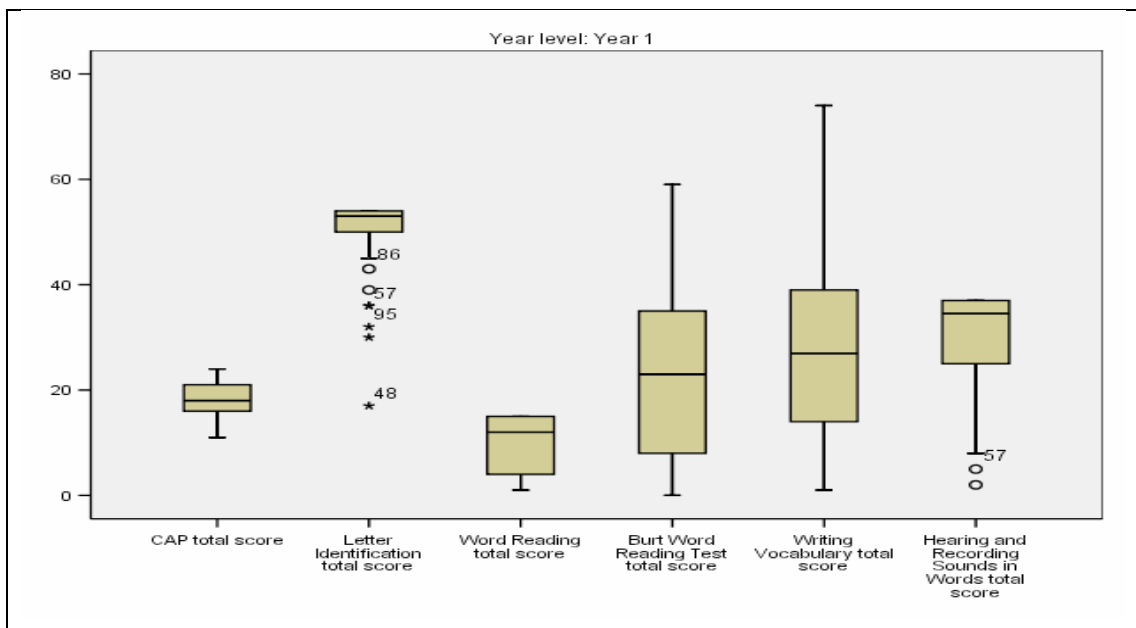


Figure 8: Test distributions for Grade 1

Figure 9 illustrates these trends even more sharply. *Concepts about Print* achievement has now moved towards the ceiling; the ceiling effect on *Letter Identification* is now clearly evident as is the (15 word) *Word Reading* total score, but the *Burt Word Reading Test* is now approximating a normal distribution as is *Writing Vocabulary*; the *Hearing and Recording Sounds* distribution has moved towards the ceiling. The utility of *Concepts about Print*, *Letter Identification*, *Word Reading* and *Hearing and Recording Sounds* is doubtful by the time students reach Grade 2.

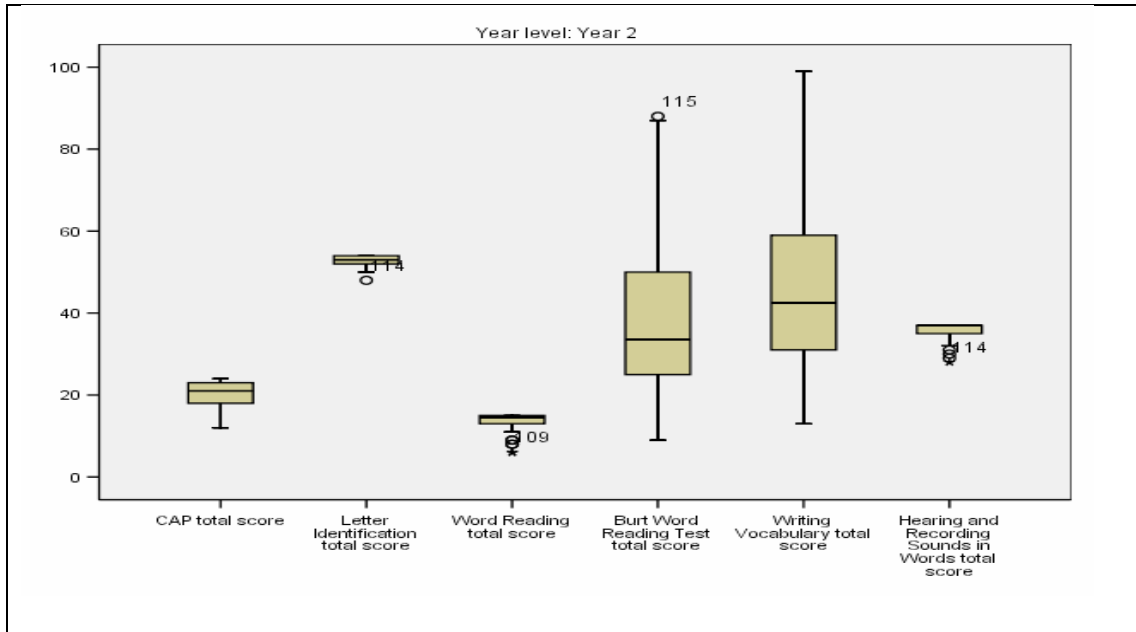


Figure 9: Test distributions for Grade 2

Table 3 presents the correlations between the six tests. They are all highly inter correlated. The lowest correlation is between the *Burt Word Reading Test* and *Letter Identification*; this could well be because *Letter Identification* quickly approaches its ceiling whereas the *Burt Word Reading Test* increases in variance with increasing age level.

Table 3: Correlations among the tests

	Letter Identification	Word Reading	Burt Reading	Writing Vocabulary	Hearing & Recording
Concepts Print Total	.75	.80	.75	.77	.83
Letter Identification Total		.70	.56	.59	.82
Word Reading Total			.87	.84	.87
Burt Reading Total				.84	.72
Writing Vocabulary Total					.76

Table 4 presents the frequencies of students administered the different booklets in the *Concepts about Print* test. It can be seen that the most popular was *Follow Me Moon*, and the least popular was *Stones*. A total of 54 students have not had recorded which booklet was used.

Table 4: Frequencies of Booklet Administrations: *Concepts about Print*

	Frequency	Percent	Valid Percent	Cumulative Percent
1 Sand	21	15.3	25.3	25.3
2 Stones	13	9.5	15.7	41.0
3 Follow Me, Moon	30	21.9	36.1	77.1
4 No Shoes	19	13.9	22.9	100.0
Total	83	60.6	100.0	
Missing	54	39.4		
Total	137	100.0		

Table 5: Frequencies of *Word Reading* Lists Administrations

	Frequency	Percent	Valid Percent	Cumulative Percent
1 List A	60	43.8	46.5	46.5
2 List B	33	24.1	25.6	72.1
3 List C	36	26.3	27.9	100.0
Total	129	94.2	100.0	
Missing	8	5.8		
Total	137	100.0		

Table 6: Frequencies of *Hearing and Recording Sounds* Forms

	Frequency	Percent	Valid Percent	Cumulative Percent
1 Form A	81	59.1	61.8	61.8
2 Form B	8	5.8	6.1	67.9
3 Form C	7	5.1	5.3	73.3
4 Form D	31	22.6	23.7	96.9
5 Form E	4	2.9	3.1	100.0
Total	131	95.6	100.0	
Missing	6	4.4		
Total	137	100.0		

Table 5 presents the data on *Word Reading* list frequencies for Lists A, B and C. It can be seen that List A was used with 60 students, almost as many as the other two lists combined. Most teachers appear to choose the first list. The *Hearing and Recording Sounds Test* has five forms, each with 37 phonemes (Table 6). The most popular of these was Form A; 81 of the 131 students were administered this form. Form D was administered to 31. These numbers are too small to break down by Grade level. By the time that we analyse Prep, Grade 1 and Grade 2, by booklet, form and wordlist, there is insufficient data available to enable the establishment of norms. What is clear from the data at each of the year levels is that the ceilings on many of

the tests are rapidly reached. This is further emphasised from results generated from item response modelling.

Item Response Analyses

The Figures in this section consist of three parts. On the left hand side there is a series of values ranging from negative to positive with zero representing the mid point of the item difficulty range. These values represent a logit scale derived from the Rasch model analysis and it provides a scale that enables a comparison of the difficulty of the items and the ability of the student. On the left hand side of the diagram is a histogram where each x represents a number of students. On the right hand side there is a series of numbers with each number representing an item. To describe the figure contents, Figure 10 reporting on *Concepts About Print*, is used as an example. Item numbers 1, 2, 3, 4, 5, 11 and 21 are missing because there was no variance in these data, or alternatively, because all the students were able to get the right answer to these questions. These items were removed from the analyses and the item numbers represented on the right hand side indicate that there is a large gap at the top of this chart. This gap is represented by an ellipse illustrating that between the most difficult item and the most able student there were no items and therefore no way to tell what the student was ready to learn.

In each of the following Figures, the mismatch between the student ability range and the difficulty range of the test items is shown by the dotted ellipse using a procedure developed by Zoanetti (2007)¹. The ellipse is placed on the variable map where there is a gap in the relationship between student ability and test item difficulty. The ellipse indicates where additional items are needed or in some cases (not here), where additional students are needed because the test is too easy or too hard and there are no students at all item levels. In many of the tests there is a need for items at both ends of the difficulty range. In some there is only a need for more difficult items. For all the tests however the caution has to be made that the numbers used for these analyses are very small and the stability of the item and student estimates is unreliable.

Another thing to notice is that the tests that are said to be equivalent, in many cases are not. The relative locations of student and item distributions across tests differ considerably. This can mean that a student's ability estimate is affected by the teachers' choice of test where there is the option of using different forms or versions.

A third feature of the variable gap maps is that the items appear to cluster in layers for some tests. If it is possible to interpret these clusters across tests it would be possible to suggest the nature and content of the proficiency levels needed to develop the skills required at the item cluster level. This in turn will help to identify the developmental levels needed to bridge to the VELS, the AIM tests and the English Continuum.

By equating the tests it would be possible to develop comprehensive levels and a development continuum based on the analyses. Series of charts presented in the

¹ Zoanetti, N. P. (2007). Technical Appendix 1. In T. N. Postlethwaite and S. Vongvichith, (Eds.), *Laos Grade 5 National Assessment Survey*.

following Figures illustrate both the ceiling effect and the lack of variance with some items.

Figure 10 represents the results for the booklets used in the *Concepts about Print*. The booklet *Stones* contained several items for which there were perfect response patterns. In this analysis it is very clear that items 1 and 3 elicit perfect responses from the group but the gap at the top is also evident where there are no items. The ceiling for the test has been reached by most students. It is also interesting to look at the variation in the order of the items for *Sand*, *Stones* and *Follow Me Moon*. Differential order indicates that the booklets might not be equivalent. Further data and analyses are needed here.

For the booklet *Follow Me Moon*, the gap at the top is less pronounced and the spread of the items is far greater than with *Stones*. What this indicates is that the booklets are not equivalent in terms of difficulty. This is further accentuated with the booklet *No Shoes* where the items are spread and the range of the ability of the students approximates the difficulty of the tasks. Each of the tests has a bimodal distribution clearly illustrating that rapid progress occurs with this skill. Each booklet however needs to be topped by developing further test items. The number of students tested with each of these booklets makes it difficult to say whether or not the booklets are equivalent or whether the items are of different difficulties within each of the booklets. Certainly, with only 19 students for *No Shoes*, 21 giving data for *Sand*, 13 for *Stones* and 30 for *Follow Me Moon*, the measurement errors would be far too great to draw firm conclusions.

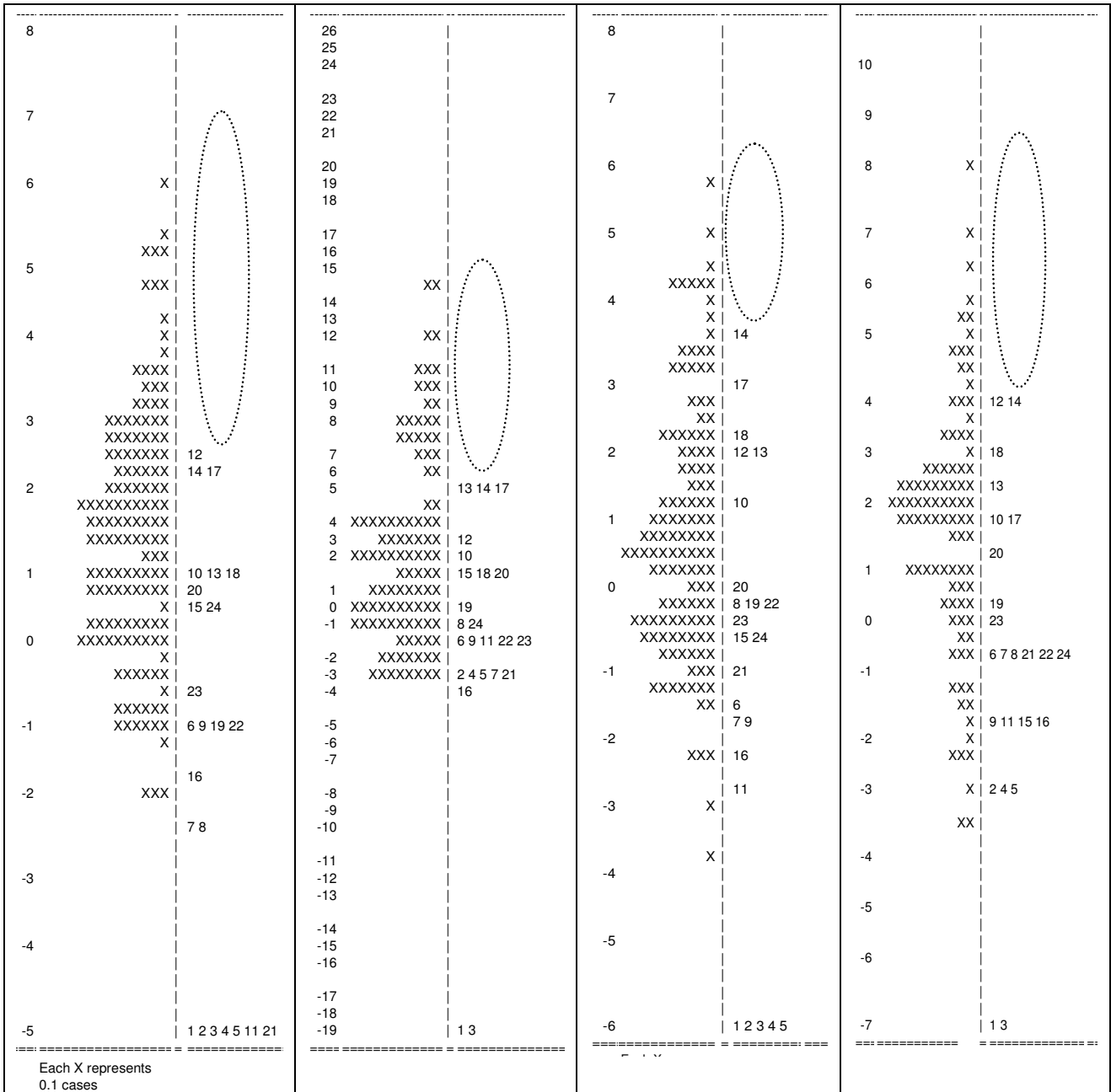


Figure 10: Variable maps for each of the books used in *Concepts about Print*
 From left to right: *Sand*, *Stones*, *Follow me*, *Moon*, and *No Shoes*

The same analyses were conducted for *Word Reading*, *Hearing and Recording Sound* and *Letter Identification*.

For the *Word Reading*, Word List A there was a very clear gap at the top and at the bottom of the lists in terms of matching the difficulty of the words to the ability of the students. The same gap could be seen with Word List B, for which 33 students presented data, and for Word List C for which 36 students presented data. These data are presented in Figure 11.

The data for *Hearing and Recording Sounds in Words* are presented in Figure 12. Form A of this test was administered to 81 students and Form D was administered to

31. There were insufficient data for Forms B and C to make any analysis. The severity of the ceiling effect in this test is very evident given the distribution of student ability versus the distribution of items in this test; the ceiling effect in Form A is clear. For Form D the ceiling effect did not appear for these 31 students and we may need to re-analyse the data with a larger sample size in order to determine the performance on this form. To be more useful, the word lists need to be topped and tailed to cater for years Prep to Grade 2.

The analysis of the *Letter Identification* test data is presented in Figure 13. It is clear that the ability of the students is far greater than the difficulty of the test but this is possibly because the ceiling is reached by the time the students have reached Grade 1. Figure 13 illustrates a clear need for extension of item difficulty on the *Letter Identification* test. For this test it may be necessary to alter the scoring procedure to allow for differing levels of performance and to top the test. Tailing the test would be difficult given the nature of the task.

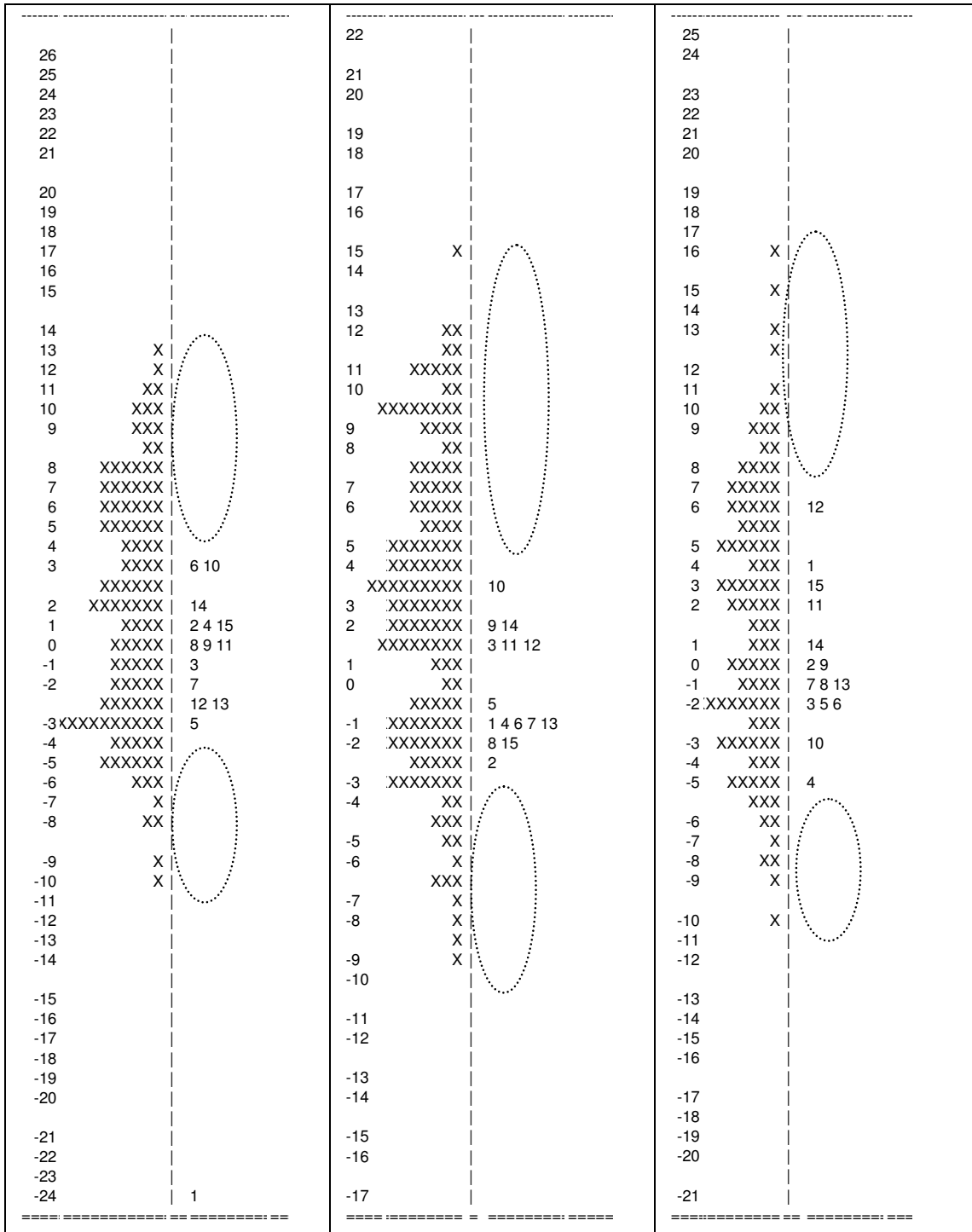


Figure 11: Word Reading Forms A, B and C from left to right

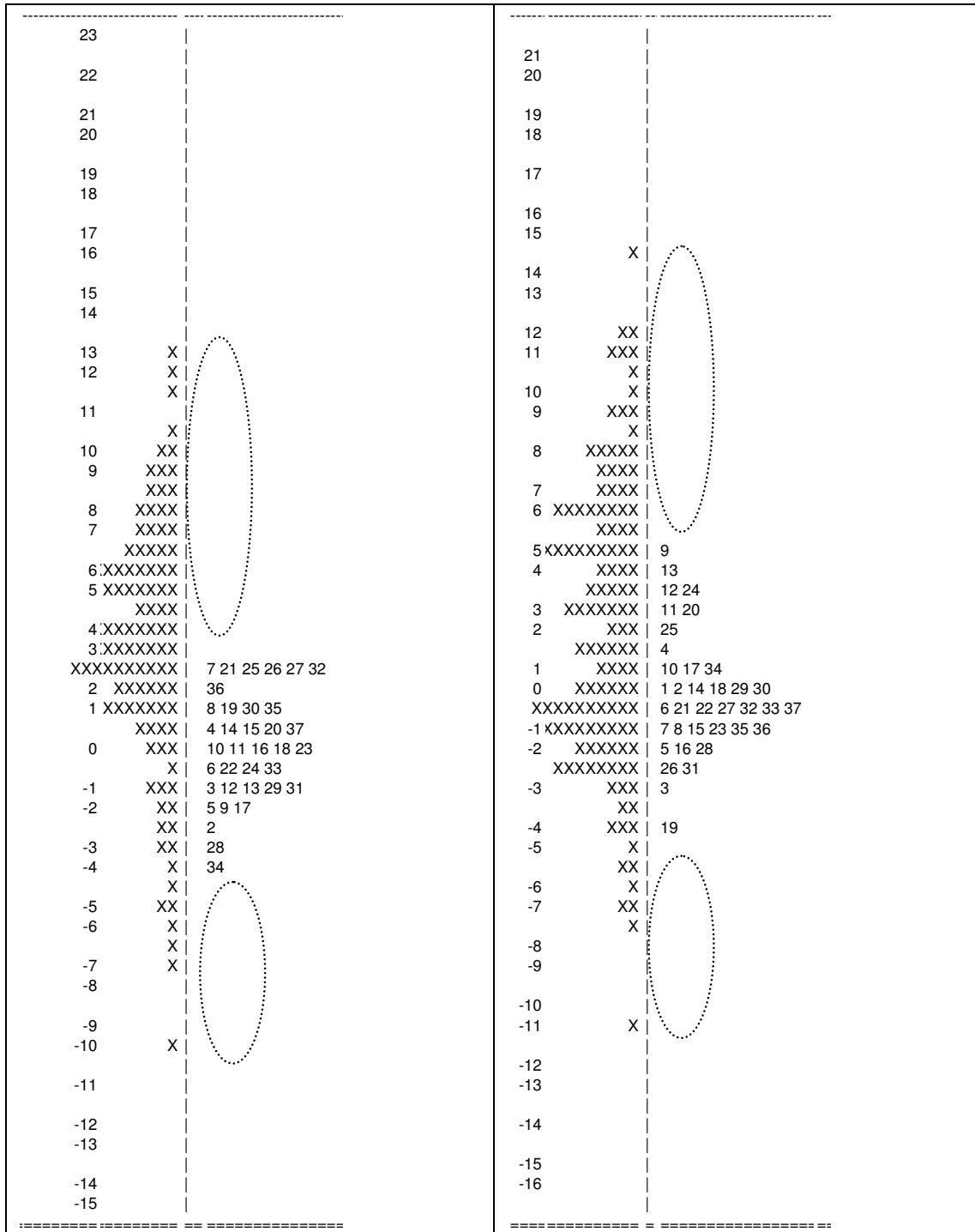


Figure 12: *Hearing and Recording Sounds* Forms A and D from left to right
 Note. Insufficient data for forms B and C.

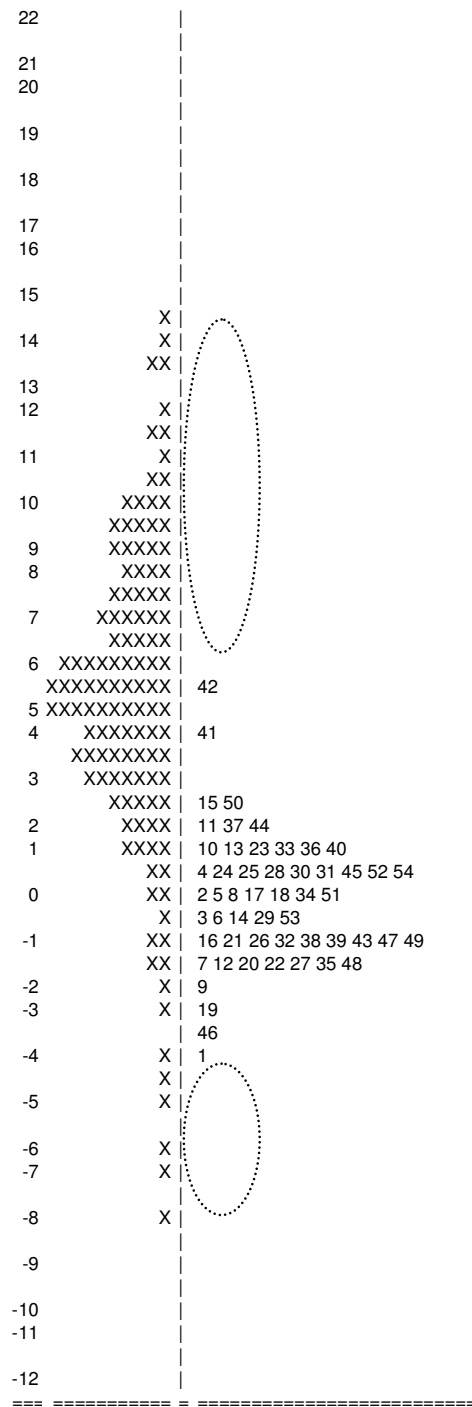


Figure 13: Letter Identification

Analysis of the *Burt Word Reading Test* is presented in Figure 14 by year level. It is very clear that there is a growth in word recognition as the median score rises almost linearly with year level from Prep to Grade 1 and to Grade 2. This might be a test that could well be administered to Grade 3 as a bridging assessment to the AIM test.

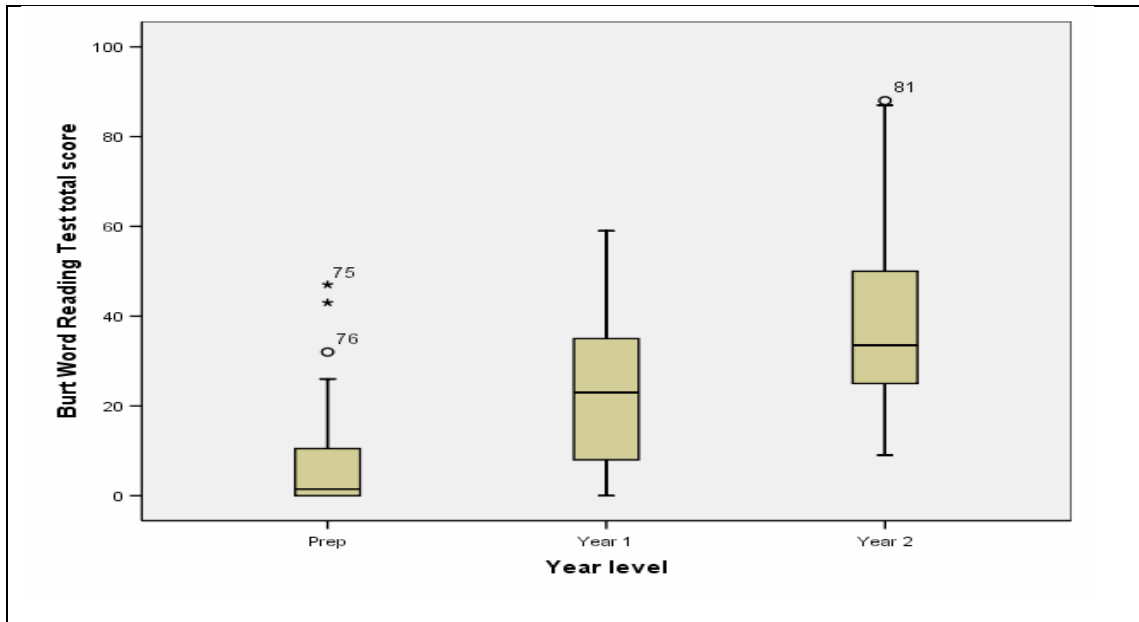


Figure 14: *Burt Word Reading Test* by year level

The Writing Vocabulary results are shown in Figure 15. There is an almost linear rise of the median and the inter-quartile range across years Prep to Grade 2.

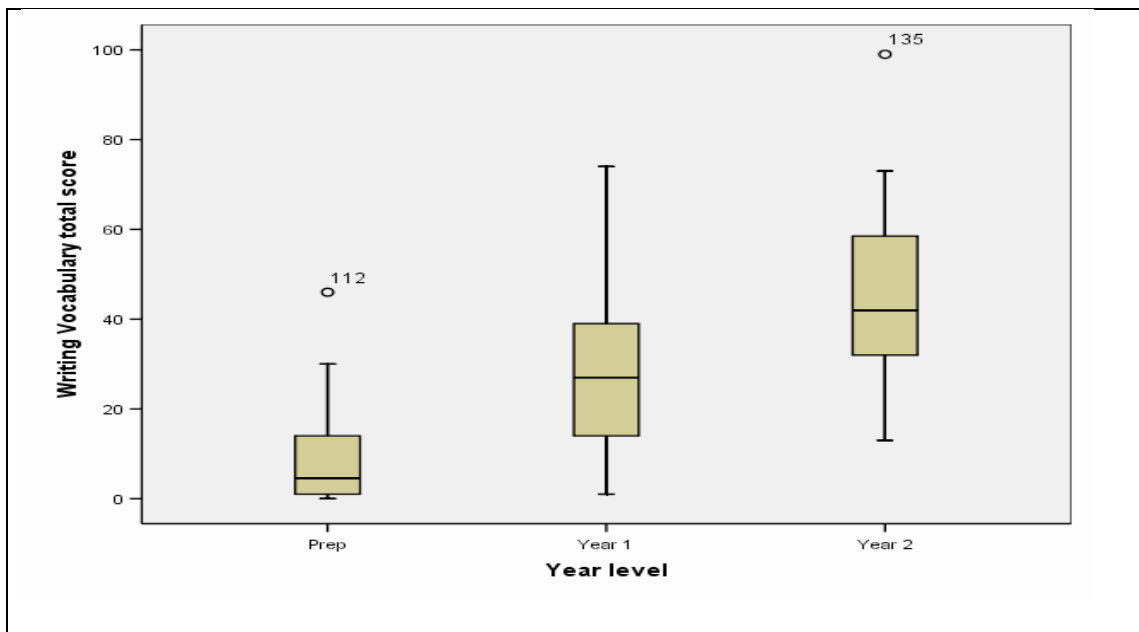


Figure 15: *Writing Vocabulary* by year level

Summary

While the data tends to show ceiling effects in three or four of the six tests, it is clear that a great deal of data is not being recorded well. It is apparent that there may be some additional training needed in test administration and scoring. Some of the criteria may not be being followed, and it appears that *Concepts about Print*, *Letter Identification*, *Word Reading* and the *Hearing and Recording* tests are inadequate beyond Prep level and would need extension if they are to be used appropriately at Grades 1 and 2.

For Grades 1 and 2 either these tests need to be extended, particularly the phoneme recognition test (*Hearing and Recording Sounds in Words*) and the *Word Reading* test. Perhaps the first and second 100 word tests could be used but the *Burt Word Reading Test* and the *Writing Vocabulary* tests could be used more extensively. The *Letter Identification* and *Word Reading* tests might be replaced by reading comprehension tests for students who have reached the ceiling in *Concepts about Print*, who have reached the ceiling in the 15 word *Word Reading* test, and who have reached the ceiling in the *Hearing and Recording Sounds* test.

Of most concern is the amount of missing data in the records. Too few students have their birth date recorded, too few have the age recorded. Under these circumstances, it is difficult to establish whether a student is developing at a rate equivalent to the published norms for the tests; it is also difficult to establish norms without that data. It means that there has to be a normative sample drawn; it would involve a very large scale of data collection to identify norms for the five forms of the *Hearing and Recording Sounds*, the three forms of the *Word Reading* and the four forms of the *Concepts about Print*. There may have to be supplementary assessments and testing in order to establish links with the AIM test for benchmarking purposes. Links to State Benchmarks should alleviate the issue where benchmarks for Prep and Grade 1 or for Grade 1 and Grade 2 appear to reach 99%, although this appears to be because the benchmarks are based on tests which are below the ability levels of the students.

Discussions need to be held with relevant personnel in the Department of Education to determine how these tests could be extended without doing a disservice to the outstanding work that Professor Marie Clay has contributed. Possibilities include the extension of *Concepts about Print* to move more to being able to retell a story from the books; the alteration of the *Letter Identification* in its scoring proforma so that it has a total score of 150 or 162 such that the student is recorded as naming, sounding and using the words; and the extension of the *Word Reading* tests beyond the *Ready to Read* series to include some of the other work, or alternatively the abandonment of the *Word Reading* test altogether, with the *Burt Word Reading Test* used in its place. It is also true that the *Hearing and Recording Sounds in Words* test needs to be extended, and a phonemic awareness and phonemic recognition test could quite readily be adopted from the work of Professor John Munro or through phonemic awareness tests that are commercially available. If these were added to the Early Years Literacy Tests the norming exercise would be able to be done, and it would also make more meaningful the links between the Early Years Literacy Tests and later Grade 3 tests of standard skills. So it appears that rather than develop totally new tests, existing instruments may be able to be used to supplement the current tests; and at least three of these can be extended beyond their current limitations.

Conclusions and Recommendations

It was clear from the analyses that the sample sizes were too small to enable the establishment of norms. A much larger sample would be needed to establish stable and useable parameters for the tests. It is also clear that many of the test forms and items are not particularly useful in the overall assessment. It is also possible that norming for year levels would be inadequate, and it is recommended that norms be established for three month age based intervals

The results from the analysis of available data needs to be treated with some caution but shows that:

- small changes are need to the instruments; larger changes are needed to their administration;
- at least one new instrument needs to be developed to bridge the gap between the AIM and the observation survey. This is recommended on the basis that it is unlikely that the AIM or the national Assessment Program will be able to accommodate the lower levels of reading comprehension and beginning writing;
- there is a ceiling effect in relation to most of the instruments;
- scoring is at times idiosyncratic, as is the matching of students to texts and tasks;
- use of alternative texts make outcomes non-comparable – using word count to determine complexity of texts would help, as would an anchor with some students using two texts;
- letter identification is useful in Prep and possibly also at Grade 1, but should be omitted from Grade 2 tests as it does not discriminate;
- there appears to be a high correlation across all instruments – this needs to be more fully explored for the different Grade and ability levels;
- reading comprehension needs to be strengthened with sentence and paragraph level texts for Grade 2.

A comprehensive approach to early literacy assessment strategies should address the following and be normed in order to enable accurate use by teachers and accurate estimates of state benchmarks.

- reading accuracy;
- reading comprehension;
- decoding (accuracy of word recognition);
- cipher knowledge (concepts about print, letter knowledge, phonemic awareness, alphabetic principles);
- lexical knowledge (vocabulary);
- linguistic knowledge (phonology, semantics, syntax).
- phonemic awareness (informed by work of John Munro);
- decoding skills
- knowledge of the alphabet (Prep and Grade 1 only)
- oral reading fluency
- knowledge of vocabulary
- reading comprehension skills
- writing skills

Instruments should encompass:

- oral reading fluency;
- vocabulary (determine text difficulty by using noun frequency, and consider using *Burt Word Reading Test* scores to decide on appropriate texts);
- strengthened reading comprehension adding sentence and paragraph level texts to Grade 2 assessment.

Instruments could comprise:

- tell me/retelling;
- record of oral language;
- observation survey (*Letter Identification, Concepts about Print, Writing Vocabulary, Hearing and Recording Sounds, Ready To Read, Running Records*).

Key actions relate to:

- removing ceilings by including more challenging items;
- strengthening items focussing on comprehension of texts (with sentence and paragraph level texts at Grade 2);
- developing criteria and guidelines, and providing professional learning to ensure consistency of administration and scoring;
- omitting letter identification from Grade 2 tests;
- monitoring correlations across instruments;
- conducting a full exercise in norming data after key actions are in place.

Assessment strategies:

- phonemic awareness
- decoding skills
- knowledge of the alphabet (Prep and Grade 1 only)
- oral reading fluency
- knowledge of vocabulary
- reading comprehension skills
- writing skills

Summary of Recommendations

It is clear from the analyses that the sample sizes were inappropriate for establishing norms. A far greater number of students are needed to establish stable and useable parameters for the tests. It is also clear that many of the test forms and items are not particularly useful in the overall assessment. It is also possible that norming for year levels would be inadequate, and it is recommended that norms be established for two or three month intervals

1. The tests *Concepts about Print, Burt Word Reading Test, and Writing Vocabulary* test should be retained – in some cases with modifications described below.
2. The *Concepts about Print* test needs to be restricted to years Prep and Grade 1. It has little or no practical advantage in being used beyond these years. The four forms of the test however need to be separately calibrated and equivalence of the skills in each aligned.

3. The *Letter Identification* test has little or no use beyond the Preparatory year in its present form. The scoring of any one or more of name, sound, or word is an inappropriate scoring protocol. We recommend that it be rescored using a three point scale for each symbol for administration at Prep level only.
4. The *Word Reading* test appears to be redundant, and should be abandoned.
5. The *Writing Vocabulary* test should be improved in terms of administration procedures, with assistance given to teachers in recording and scoring. Also, students should be asked to write short sentences for this test once they reach Grade 1. Scoring rubrics and administration protocols are needed for this writing component.
6. The *Hearing and Recording Sounds* test needs to be extended; the 37 phonemes are not equivalent in the four forms and the ceiling is reached very early. It is recommended that both phoneme awareness and phoneme hearing (or auditory processing) be assessed. The tests as they stand may be confounding these two developmental skills.
7. A new test does appear to be warranted. It should emphasise the *bridge* needed between the most demanding comprehension skills in the Observation Survey and the easiest or least demanding skills needed in the Year 3 AIM test. The test should consist of sentence length text to be used as a basis for assessing listening and reading skills through both written and physical responses (e.g., pointing) from the students. The test should also contain items such as those found at the easier end of the AIM test. Two forms of the test appear to be needed. Form A should be a short sentence and word picture test administered individually to the student who would be requested to point or provide oral answers. Form B should be a group administered test administered to students from 7 years of age; it would require written answers. During the development phase, it would need to be administered to a sample of children from ages 7 to 9 and if possible also administered to a sub sample of students who would also complete the Grade 3 AIM test.
8. As can be seen from the Figure 16 a sample size of 4800 would be needed to norm the Observation Survey using age based samples targeted at quarter years. The administration of components would also be administered to sub samples as shown, as would the new *bridge* test and the extended *Hearing and Recording Sounds* in Words test.
9. A concerted effort is needed to provide refresher in service training for teachers in the administration, scoring and interpretation of the Observation Survey and its extension.

Concepts about Print	Form	Age in years / quarters																
		5.1	5.2	5.3	5.4	6.1	6.2	6.3	6.4	7.1	7.2	7.3	7.4	8.1	8.2	8.3	8.4	
	A	300	300	300	300	300	300	300	300	300	300	300						
	B	300	300	300	300	300	300	300	300	300	300	300						
	C	300	300	300	300	300	300	300	300	300	300	300						
	D	300	300	300	300	300	300	300	300	300	300	300						
Letter Identification																		
Name, Sound, Word		300	300	300	300	300	300	300	300	300	300	300	300	300				
The Burt Word Reading Test		300	300	300	300	300	300	300	300	300	300	300	300	300				
Writing Vocabulary		300	300	300	300	300	300	300	300	300	300	300	300	300	300	300		
Writing sentences																		
Hearing and Recording Sounds in Words																		
	A	300	300	300	300	300	300	300	300	300	300	300						
	B	300	300	300	300	300	300	300	300	300	300	300						
	C	300	300	300	300	300	300	300	300	300	300	300						
	D	300	300	300	300	300	300	300	300	300	300	300						
Extended Form	E												300	300	300	300	300	
AIM Bridge Test																		
Sentence and picture	A												300	300	300			
Read and record	B												300	300	300	300	300	
AIM test													300	300	300	300	300	
Sample size		300	300	300	300	300	300	300	300	300	300	300	300	300	300	300	300	4800

Figure 16: Action summary for establishing norms and benchmark assessments.

Sample Sizes

Any exercise in establishing norms need to ensure that the data are stable and that measurement and sampling errors are minimised. Population figures would be required for correct sampling estimates.

However as a first estimate, approximately 1200 students might be systematically but scientifically sampled per grade level sampled using age as a weighing variable. This would enable norms to be estimated for developmental age groups over the three years from Prep to Year 2.

For the re-norming of the Observation Survey, a preliminary and minimum sample size per age group of 300 students is recommended. This overall number of test items will make the test reliable and stable.

Age based samples are needed with approximately 200 if the age groups are 2 month or 400 if it is decided to use 3 month. These sample sizes will yield suitably small sampling errors for norming.

An overall sample of 4800 students would be needed over the years Prep to Grade 3.

Reference

Clay, M. M. (2002). *An Observation Survey of Early Literacy Achievement*. NZ: Heinemann.