Pilot study to develop a new version of the Hearing and Recording Sounds in Words task

Darr, C., Lawes, E., Mazengarb, J.,
Eyre, J., and McDowall, S.

NZCER
Rangahau Mātauranga o Aotearoa

2019

Introduction

The Marie Clay Literacy Trust (MCLT) contracted the New Zealand Council for Educational Research (NZCER) to develop a new version of the task Hearing and Recording Sounds in Words. This task is one of six in An Observation Survey of Early Literacy Achievement (Clay, 2013).

The Hearing and Recording Sounds in Words task is designed to provide information about what a child knows about phonemic awareness and sound–letter (phoneme–grapheme) relationships. These are known to be important underpinning skills for early literacy development.[1]

The Marie Clay Literacy Trust asked NZCER to revise the task with the aim of reducing the ceiling effect that is observed when the task is used with children who are more able readers and writers. The Trust asked NZCER to develop five new forms for the task, each of equivalent difficulty.

This report describes the work undertaken by NZCER to develop the new forms. The report describes the methodology used and outlines the results of a pilot study involving the new forms. It is intended that the results of the pilot study are used to guide which forms are finalised for use in a larger norming study in 2018.

The Hearing and Recording Sounds in Words task

The Hearing and Recording Sounds in Words task assesses phonemic awareness and knowledge of sound–letter relationships in combination. A series of words, in the form of one or two short sentences, is read out to the child, word by word. The child records the sounds they hear in each word.

The task is scored by noting each sound (phoneme) the child has recorded correctly.

In the current version of the task there are five short passages or 'forms' that the observer can select from to read aloud to the child. Each form consists of at least one sentence and has a maximum score of 37 correctly recorded phonemes.

The current forms are presented in Table 1.

Table 1 Current forms for the current Hearing and Recording Sounds in Words task

Form A	I have a big dog at home. Today I am going to take him to school.
Form B	Mum/Mom 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. He can go very fast on it.

Each of the current forms includes a subset of the 44 possible phonemes. Some phonemes are common to all the forms (e.g., /b/) and some phonemes occur in only one form (e.g., /sh/ in form B). Therefore, there are some similarities and some differences between the phonemes assessed by each form.

A ceiling effect can be observed when the current five forms are administered to students in the higher achievement range. For instance, for students in the 6.51 to 7.00 years age-group the average score is 36 out of 37. (For more information about the ceiling effect see the stanine tables for the current forms in Appendix 3.)

1 See, for example; Clay, 2014, Foulin, 2005, Goswami, 2008, Kim, Petscher, Foorman, and Zhou, 2010

Methodology

NZCER began the development process by finding out how the existing forms were constructed, including the theory that informed their development. Material relevant to the task was located and conversations held with members of the National Reading Recovery team. NZCER also located and read research literature relevant to the assessment of hearing and recording sounds (i.e., phonemic awareness and knowledge of phoneme–grapheme relationships). The findings from the scoping exercise were used to write a brief rationale for the proposed approach to augmenting the Hearing and Recording Sounds in Words task, which was shared and discussed with the MCLT.

NZCER's next step was to construct seven new forms for the Hearing and Recording Sounds in Words task. The process involved identifying a key set of criteria for what would be included in each of the new forms and then using these criteria to construct sentences and words for the new forms. To increase the demand, more phonemes were included in each new form compared with the current forms. Each new form included some one or two-phoneme words near the beginning of the sentence to build confidence (e.g., 'I', 'a', at); at least one word with an initial consonant blend (e.g., 'growls', 'sprained'); at least one word with a two-letter phoneme (e.g., 'cheese', 'shampoo'); and at least one word with six or more phonemes (e.g., 'splashing', 'delicious').

The new forms were piloted informally with a small group of students in local Wellington schools and adjustments made as necessary.

NZCER worked with National Reading Recovery to invite all Tutors from around New Zealand to administer the new forms in the school where they normally teach individual children. Those that agreed to take part were asked to discuss the project with the Principal at the school to gain their approval to trial two of the new forms with up to 12 students from the school.

The Tutors were asked to select three students aged 5:00 to 5:50 years, three aged 5:51-6:00 years, three aged 6.01-6.50 years, and three aged 6.51-7.00 years from classes in that school.

Each student completed two of the new forms in a single session. An administration design was applied that ensured each form was used with students in all of the various age categories, was variously paired with each of the other forms involved in the trial, and was administered as both the first and second form.

Tutors administered the new forms using the protocols described in *An Observation Survey of Early Literacy Achievement* 3rd ed., p.118.

Scoring was carried out by two researchers at NZCER using the same scoring protocols as for the current forms. Cross-marking was used to check that the scoring protocols were applied consistently.

For each student the number of phonemes recorded correctly for each scored word was entered into a data file for both of the forms administered.

The data were analysed using the partial credit form of the Rasch model (Masters, 1982). The administration design allowed data from each form to be analysed together in one joint analysis. This allowed the difficulty of each form to be compared using the same underlying scale.

Pilot forms for the revised Hearing and Recording Sounds in Words task

Table 2 shows the seven forms (Forms A to G) that were piloted. For each form the number of phonemes scored per word is shown.

Table 2 The piloted Hearing and Recording Sounds in Words forms

Form	Passage
A	My puppy likes it when I scratch her tummy. Sometimes she growls at my friends. 2 4 4 2 3 1 5 3 4 7 2 5 2 0 6
B	At the supermarket, we bought bread, milk, tea, shampoo, and strawberry ice cream. 2 2 9 2 3 4 4 2 5 3 8 2 4
C	I made popcorn in a big pot. It was crunchy and sweet and tasted delicious. 1 3 6 2 1 3 3 2 3 6 3 4 0 6 7
D	For my lunch I have a crunchy apple, a slice of bread, two carrots, and a piece of cheese. 2 2 4 1 3 1 6 4 0 4 2 4 2 6 3 0 3 0 3
E	I tripped and sprained my ankle yesterday. My teacher wrapped a bandage round it. 1 5 3 6 2 5 7 0 4 4 1 6 4 2
F	At the zoo the three animals I like best are chimpanzees, bears, and elephants. 2 2 2 0 3 7 1 3 4 2 9 4 3 8
G	My baby sister likes playing in the bath. She splashes and makes the floor wet. 2 4 5 4 5 2 2 3 2 7 3 4 0 4 3

Notes: a. words repeated in a form were not scored on second or subsequent occasions
b. the r-controlled vowels 'or' in 'floor', 'ear' in 'bears', and 'er' in 'her' were scored as two phonemes.

Results

The students

In total, 258 students were assessed using the new forms. The students came from 22 schools. Table 3 shows the number of students according to their school decile rating. The deciles have been grouped into low (deciles 1 to 3), mid (deciles 4 to 7) and high (deciles 8 to 10).

Table 3 Number of schools by decile grouping

Decile Group	Low (1-3)	Mid (4-7)	High (8-10)
Number of schools	10	9	3

Table 4 shows the number of students who completed each form by age group along with the total number of students from each age group involved in the pilot. The age groups used are the same as those applied to create norms for the current forms.

Table 4 Number of students completing each form by age group

Form	5:00-5:50	5:51-6:00	6:01-6:50	6:51-7:00
A	18	20	18	18
B	18	24	18	16
C	20	21	14	18
D	18	21	14	20
E	22	23	14	15
F	18	19	23	12
G	18	15	17	23
All forms	66	72	59	61

Summary statistics by form

Table 5 provides summary statistics based on the raw scores for each form. Overall the average scores were similar, ranging from 30.0 to 36.0.

Table 5 Summary statistics for each form based on raw scores

Form	Number of students	Maximum possible score	Mean	Standard deviation
A	74	50	34.7	14.6
B	76	50	35.0	13.5
C	73	50	34.8	12.9
D	73	50	35.7	14.3
E	74	50	31.4	14.5
F	72	50	30.4	14.8
G	73	50	36.0	14.8

Internal reliability

Cronbach's Alpha was calculated for each form as an index of internal reliability. Table 6 shows the Cronbach's Alpha by form. These were consistently high, ranging from 0.93 to 0.96.

Table 6 Cronbach's Alpha by form

Form	Cronbach alpha
A	0.95
B	0.94
C	0.93
D	0.96
E	0.95
F	0.93
G	0.95

Form equivalence

The first part of the analysis involved applying the partial credit form of the Rasch model (Masters, 1982) to the data associated with each of the forms separately.

The Rasch model is a mathematical model concerned with describing the probabilistic relationship between a student's overall ability and the different levels of response associated with test items. The model estimates where each student's ability is located on a measurement continuum. It also locates an estimate of the demand (difficulty) associated with different levels of response for each item on the same continuum.

For this analysis the test items were defined as the scored words in each form (see Table 2). Form A, for example, was made up of 14 items. For each item the level of response was defined as the number of phonemes recorded correctly. This meant, for example, that for Form A the level of response to the word 'I' (one phoneme) was coded as either '0' or '1', while responses to the word 'her' (three phonemes) were coded as either '0', '1', '2', or '3'.

Statistical and graphical fit indices were used to examine the extent to which the data associated with each form met the assumptions of the Rasch model. That is, how well did the Rasch model predict the level of response on each word, given a student's overall ability level? The fit indices indicated that, overall, the data for each form showed good fit to the model.

A summary of the analysis associated with each form is provided in Appendix 1. Each summary includes an item map that locates both the Rasch estimates of student ability and the relative demand (difficulty) associated with each item on the measurement continuum generated by the analysis. Figure 1 below shows the item map associated with Form A. The figure shows the measurement continuum (scale) on the left-hand side labelled in logits (the probabilistic measurement unit used by the Rasch model) from -6 to 8. Each of the words (items) that were scored in Form A are located on the right of the continuum. Words that are located higher on the continuum were found to be relatively more difficult overall than those located at lower levels. The most demanding word overall in Form A was 'growls'. The word 'I', on the other hand, was relatively easy.

The 'X's to the left of the vertical line represent the ability locations associated with the students who completed Form A. Students with ability estimates located higher on the measurement scale are more likely to have correctly recorded the phonemes associated with each word than those with ability estimates located lower on the scale. The map indicates that there are a number of students with total ability locations that are well above the location of the most difficult word. This indicates that many students were highly successful in dealing with the challenges presented by the form.

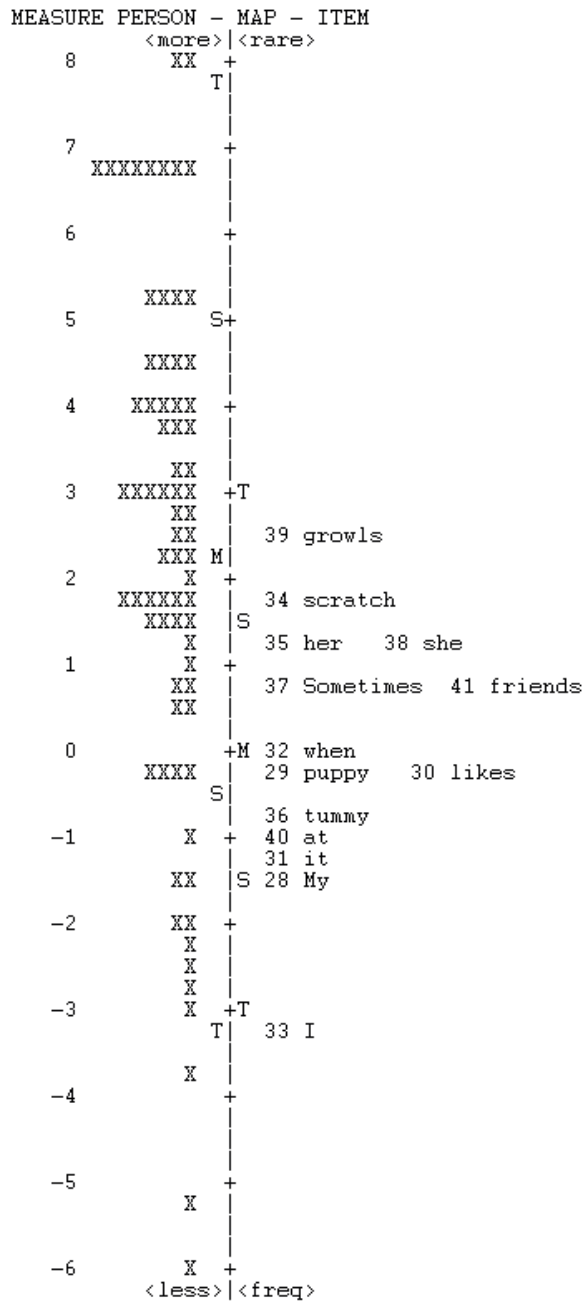


Figure 1 The item map for Form A

To study the extent to which the forms could be considered to be equivalent, a joint analysis was carried out. This involved combining data from all the forms into one analysis. Responses to each form were linked by the fact that each student had done two forms and that various pairings of forms had been allocated to different students. The joint analysis allowed the relative difficulty of the items for each form to be located on the same measurement scale and thus the difficulty of the forms to be compared.

Table 7 shows the mean and standard deviation of the difficulty thresholds associated with each word in each form. Overall the forms were of similar difficulty with the mean threshold locations ranging from 0.32 logits (Form D) to 0.85 logits (Form E) — a range of about 0.5 logits.

The standard deviations for the forms varied from 1.523 logits (Form G) to 2.51 logits (Form F).

Table 7 Average item threshold difficulty and standard deviations by form

Form	Mean item threshold difficulty (logits)	Standard deviation (logits)
A	0.41	2.26
B	0.54	2.10
C	0.56	1.83
D	0.32	1.98
E	0.85	2.24
F	0.72	2.51
G	0.63	1.52

Test equivalency was also considered by plotting test characteristic curves (TCCs) for each of the forms. A TCC shows the expected scores on a particular test across an ability continuum. When two different test forms are equivalent, the TCCs for the forms should overlap.

Figure 2 shows a plot of the TCCs for the seven different forms. Overall there was substantial overlap between the curves. However, there was some variation, with the TCCs for forms E and F generally positioned to the right of the TCCs for the other forms. This indicates that, in general, on these forms students require higher overall ability to achieve a given raw score compared with achieving the same score on one of the remaining forms.

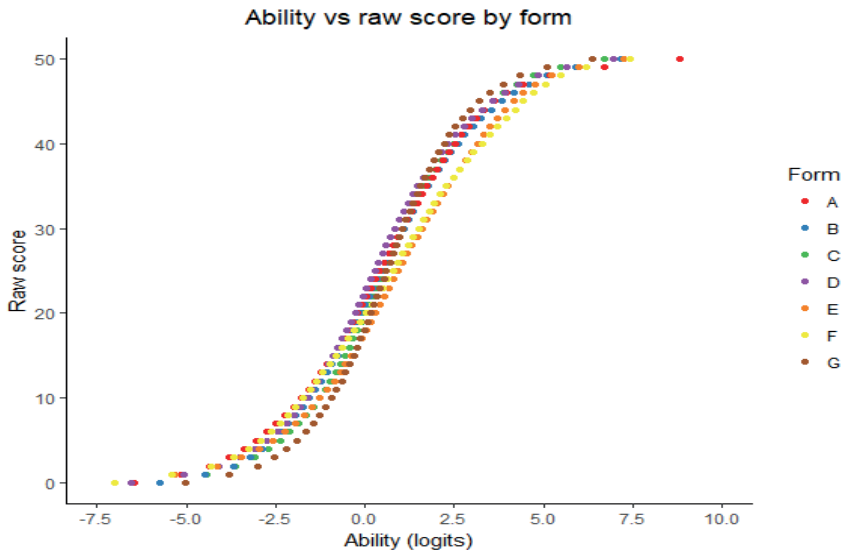


Figure 2 Test Characteristic Curves for Forms A, B, C, D, E, F and G

Figure 3 shows how the TCCs overlap after the curves for Forms E and F are removed. The remaining TCCs show fairly close alignment across the ability continuum.

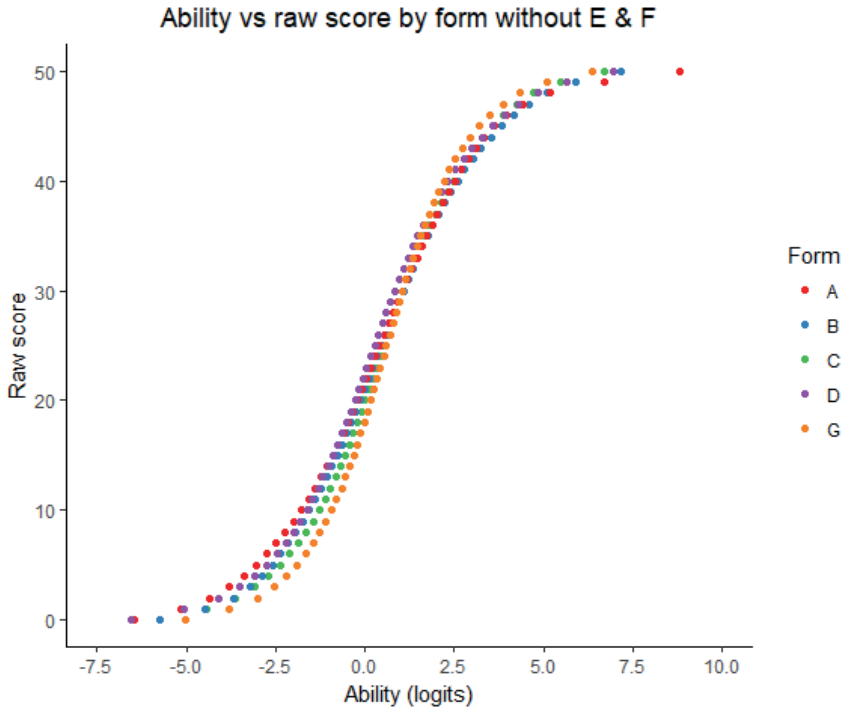


Figure 3 Test Characteristic Curves for Forms A, B, C, D and G

Figure 4 goes one step further by also removing the curve for Form G.

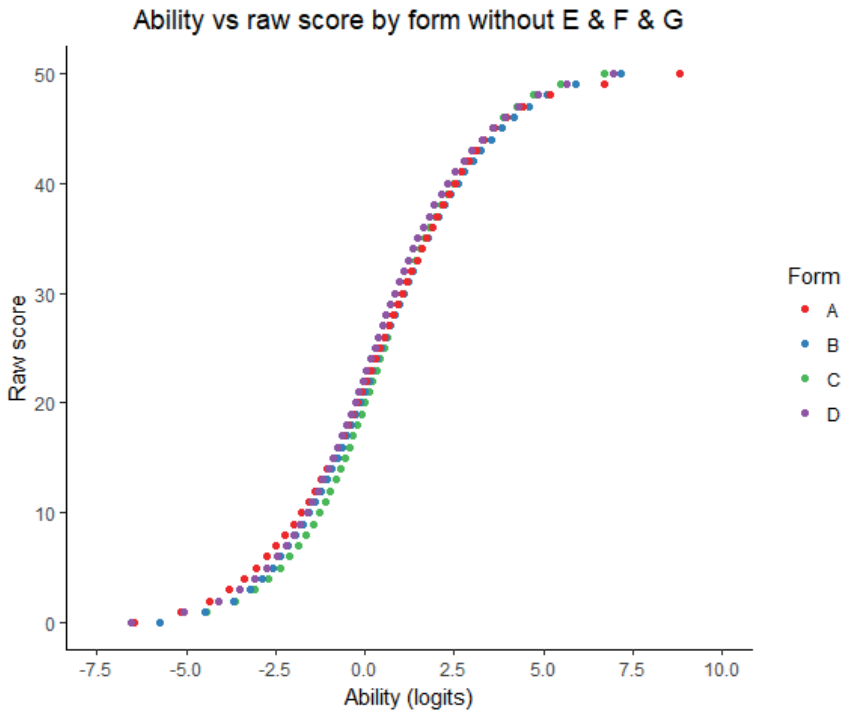


Figure 4 Test Characteristic Curves for Forms A, B, C, and D

Table 8 considers the impact of differences between the forms in terms of raw score to stanine conversion. This is done by comparing the stanine conversions for Forms C, D and E[1]. The stanine conversions are based on normalised distributions of student achievement based on the mean and standard deviation of scores for all students aged 5:00 to 5:50 years who were involved in the pilot.

Table 8 Raw Score to Stanine conversion table for Forms C, D and E: 5:00-5:50 years

Stanine	1	2	3	4	5	6	7	8	9
Form C	0	1-2	3-4	5-9	10-18	19-29	30-38	39-44	45-50
Form D	0-1	2	3-6	7-12	13-21	22-31	32-39	40-44	45-50
Form E	0-1	2	3-5	6-9	10-16	17-25	26-34	35-41	42-50

Overall the conversions between raw score and stanine are similar across the forms. The score ranges associated with each of the stanine categories overlap and any difference between forms in the conversion between raw score and stanine is never greater than one stanine.

Equivalency coefficients

Parallel forms are sometimes compared by calculating an equivalency coefficient. This is simply the correlation between two sets of scores for a group of students who have taken both tests. A form of equivalency coefficient for each of the piloted forms was calculated by correlating the scores on each form with the scores achieved by the same students on the remaining forms (see Table 9). These were all very high, ranging from 0.91 to 0.96.

Table 9 Correlations between the scores on each form and scores on the remaining forms

Form	Correlation with scores on alternate forms (Pearson correlation)
A	0.92
B	0.95
C	0.94
D	0.95
E	0.95
F	0.96
G	0.91

Examining the ceiling effect

The ability of the forms to accommodate the ability distributions of students in the older age category was explored by constructing stanine tables for each form based on the 6:01 to 6:50 and 6:51 to 7:00 age groups. The tables associated with Form C are provided below[2]. Also provided for comparative purposes are the respective stanine conversion tables for the current forms. As above, the ability distributions were based on normalised distributions of student achievement calculated using the mean and standard deviation of all students in these age categories that were involved in the pilot study. Although the age group samples are relatively small and not necessarily nationally representative, they are drawn from students in a range of schools covering the full range of deciles. It is important to note that there were fewer high decile schools compared to low decile schools in the sample.

1 These forms were selected to cover the range of difficulty associated with the different forms.

2 Form C is chosen as a typical example of one of the new forms.

Each conversion table is presented alongside the matching stanine conversion table for the current forms. The tables indicate that the new forms do exhibit less of a ceiling effect compared with the original forms. However, the new forms do still appear to struggle to differentiate high achievement in the older age group category.

Table 10 Stanine conversion table for pilot Form A: 6:01-6:50 years

Test score	0-22	23-29	30-35	36-40	41-43	44-46	47	48	49-50
Stanine	1	2	3	4	5	6	7	8	9

Table 11 Stanine conversion table for the current forms: 6:01-6:50 years

Test score	0-8	9-19	20-27	28-32	33-35	36	37		
Stanine	1	2	3	4	5	6	7	8	9

Table 12 Stanine conversion table for pilot Form A: 6:51-7:00 years

Test score	0-29	30-36	37-41	42-45	46-47	48-49	50		
Stanine	1	2	3	4	5	6	7	8	9

Table 13 Stanine conversion table for the current forms: 6:51-7:00 years

Test score	0-14	15-28	29-32	33-35	36	37			
Stanine	1	2	3	4	5	6	7	8	9

Concluding comments

Overall the pilot study indicates that the new forms are working well to measure the competencies assessed by the Hearing and Recording Sounds in Words task. They all have strong internal reliability and responses to the forms show strong fit to the Rasch model. The forms also appear to be of similar difficulty. The Rasch analysis indicates that they function in a similar way across an ability continuum and that they produce similar stanine conversion tables. The forms also appear to differentiate more clearly between high achieving students than is the case for the current forms. However, the new forms do still appear to struggle to differentiate between high achievers in the older age group category (6:51 to 7:00 years).

The data from the pilot study indicates that students, overall, make a large amount of progress across the age groups. The Cohens' D effect size related to the difference in average scores between the oldest and youngest age groups in the pilot was 2.52. This indicates that it would be difficult to develop a single assessment form that can cope with the range of abilities based on the current format. If, for instance, the forms were made longer and more difficult words included, the forms could become less useful at the younger age levels.

Recommendations

NZCER recommends that five forms from the pilot study are prepared for the norming study. We recommend sentences A, B, C, D, and G because these showed the greatest similarity in level of difficulty.

A My puppy likes it when I scratch her tummy. Sometimes she growls at my friends.
B At the supermarket, we bought bread, milk, tea, shampoo, and strawberry ice cream.
C I made popcorn in a big pot. It was crunchy and sweet and tasted delicious.
D For my lunch I have a crunchy apple, a slice of bread, two carrots, and a piece of cheese.
G My baby sister likes playing in the bath. She splashes and makes the floor wet.

We are aware that two of the international reviewers expressed concern that children in Canada and the USA would not be familiar with the concept of popcorn being sweet (Form C). Ideally, forms would contain ideas and experiences that most children are familiar with. However, there will always be some children that are unfamiliar with some of the words and concepts used. It is important to remember that the forms are not designed to assess children's world knowledge, but their ability to hear and record sounds in familiar and unfamiliar words. We consider the words and concepts of 'sweet' and 'salty' will be within most children's experience. So even if a child has not experienced sweet popcorn, we don't think this will adversely affect their ability to hear and record the sounds in the word 'sweet'. Further, teachers will have a choice of which forms will best suit their children.

Appendix 1: Rasch analyses by form

Form A

Table 14 Rasch item measures and corresponding Infit Mean Square statistics for Form A

Item	Measure (logits)	Infit MNSQ[4]
28	-1.46	1.01
29	-0.26	0.66
30	-0.19	0.92
31	-1.15	0.99
32	0.00	0.93
33	-3.37	0.98
34	1.77	0.73
35	1.15	0.83
36	-0.84	0.64
37	0.84	0.75
38	1.18	1.22
39	2.60	0.94
40	-1.11	1.06
41	0.64	1.32

4 The Infit Mean Square statistic is a fit indicator used to check whether a test item is meeting the assumptions of the Rasch model. A value close to 1indicates strong fit.

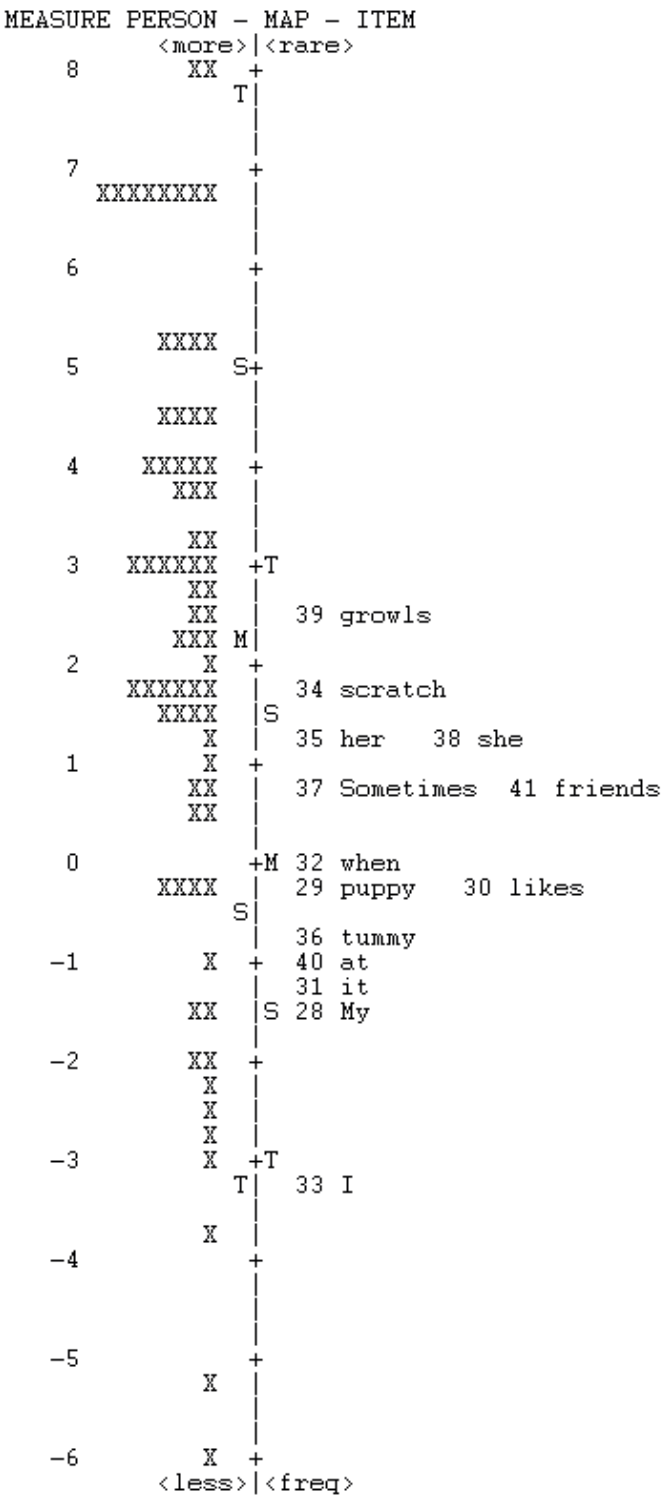


Figure 5 Item map for Form A

Form B

Table 15 Rasch item measures and corresponding Infit Mean Square statistics for Form B

Item	Measure (logits)	Infit MNSQ
15	-1.38	0.85
16	-2.08	2.65
17	1.19	1.21
18	-1.68	1.25
19	0.14	0.91
20	0.57	0.64
21	0.41	0.85
22	-1.26	0.85
23	1.29	0.84
24	-.18	0.91
25	2.21	0.88
26	-0.61	1.50
27	0.64	0.76

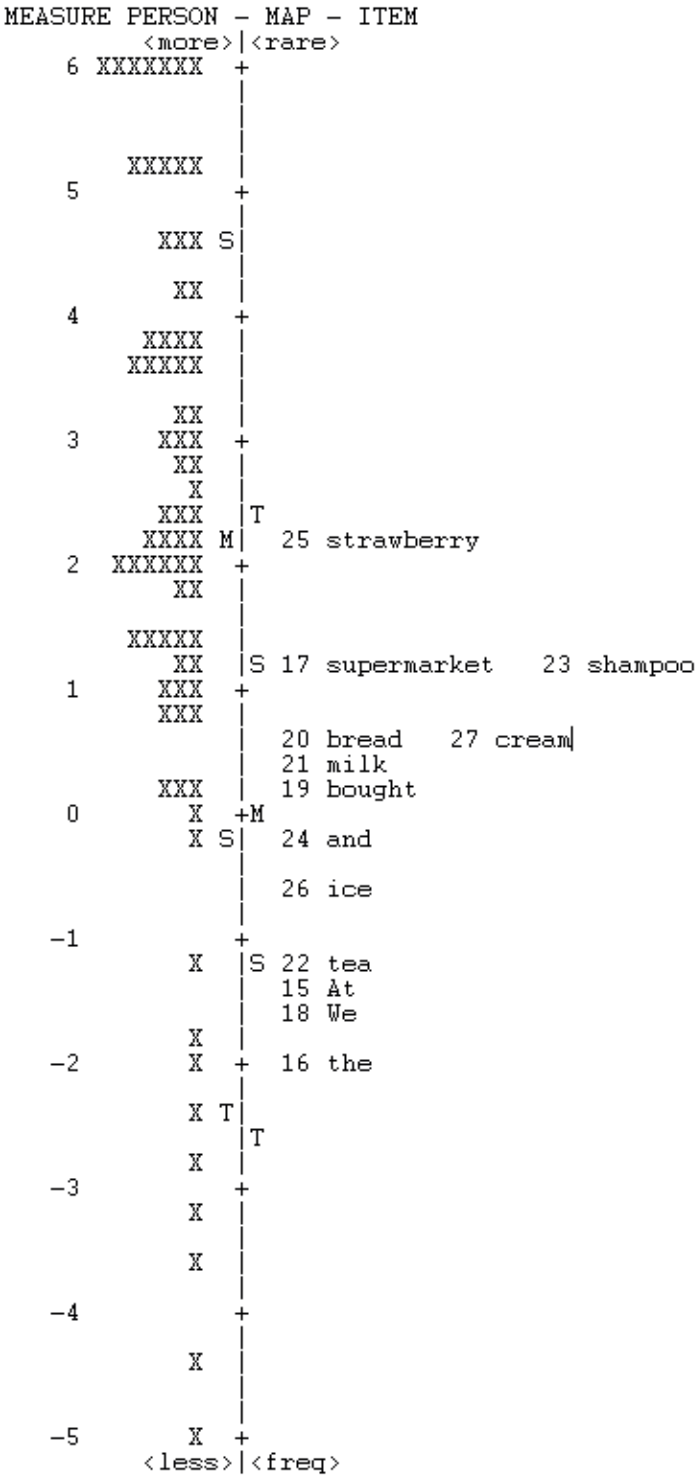


Figure 6 Item map for Form B

Form C

Table 16 Rasch item measures and corresponding Infit Mean Square statistics for Form C

Item	Measure (logits)	Infit MNSQ
1	-3.37	0.27
2	-0.44	0.73
3	0.78	0.77
4	-0.94	0.89
5	-1.90	0.71
6	0.20	1.21
7	-0.42	0.73
8	-0.90	0.95
9	-0.12	0.64
10	1.77	0.79
11	0.03	1.61
12	0.63	1.38
13	1.59	1.34
14	1.78	1.02

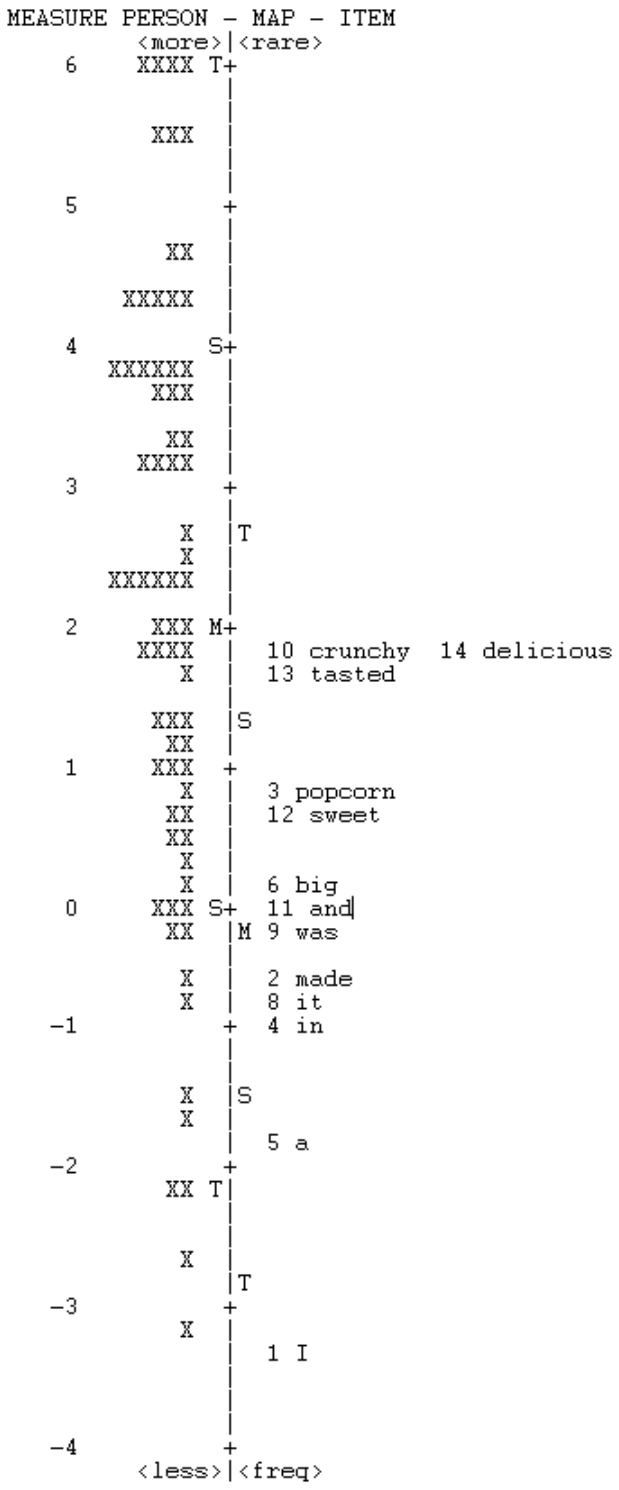


Figure 7 Item map for Form C

Form D

Table 17 Rasch item measures and corresponding Infit Mean Square statistics for Form D

Item	Measure (logits)	Infit MNSQ
42	-0.11	1.23
43	-1.06	1.66
44	1.34	0.91
45	-5.05	1.23
46	0.41	0.68
47	-1.35	0.71
48	1.92	0.53
49	-0.16	0.52
50	0.53	0.97
51	-1.50	0.75
52	0.71	0.58
53	-0.91	1.65
54	0.84	0.72
55	-0.29	0.90
56	-0.01	0.59
57	1.21	1.23

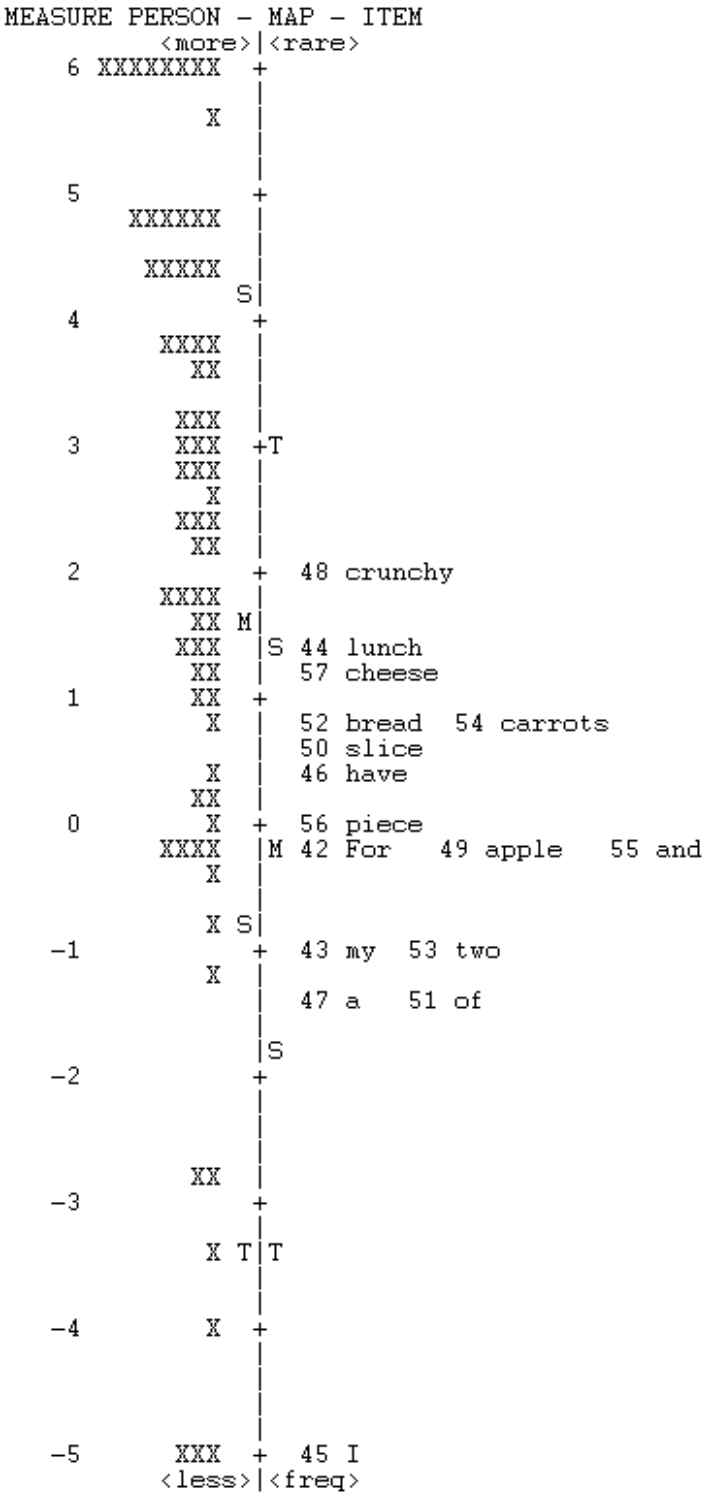


Figure 8 Item map for Form D

Form E

Table 18 Rasch item measures and corresponding Infit Mean Square statistics for Form E

Item	Measure (logits)	Infit MNSQ
58	-5.81	1.01
59	0.99	1.39
60	-0.01	1.18
61	1.52	0.63
62	-1.64	1.31
63	1.22	0.79
64	1.41	0.79
65	1.21	0.81
66	0.49	0.53
67	-1.99	0.60
68	1.75	0.91
69	2.17	0.57
70	-0.76	0.90

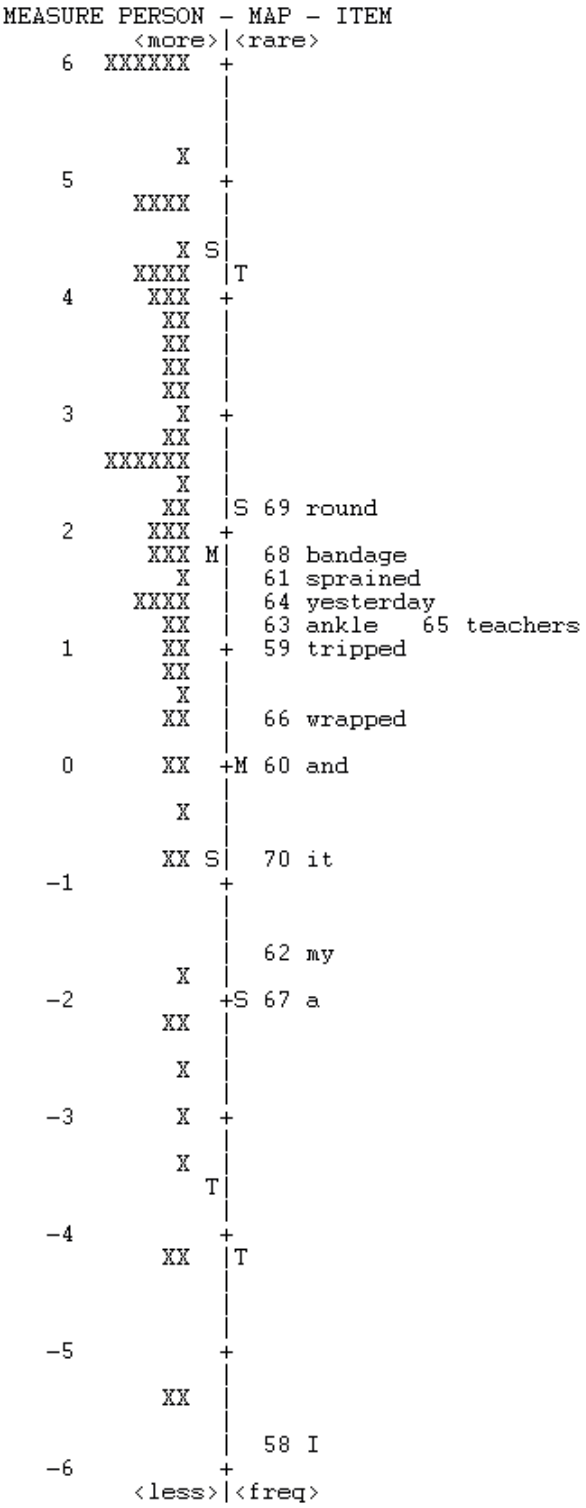


Figure 9 Item map for Form E

Form F

Table 19 Rasch item measures and corresponding Infit Mean Square statistics for Form F

Item	Measure (logits)	Infit MNSQ
71	-1.21	1.03
72	-2.06	1.20
73	-0.99	1.58
74	2.39	1.41
75	1.28	0.88
76	-5.68	1.25
77	-0.95	1.14
78	0.05	0.89
79	-0.05	0.98
80	2.45	0.84
81	0.46	0.86
82	-0.07	0.87
83	1.42	0.78

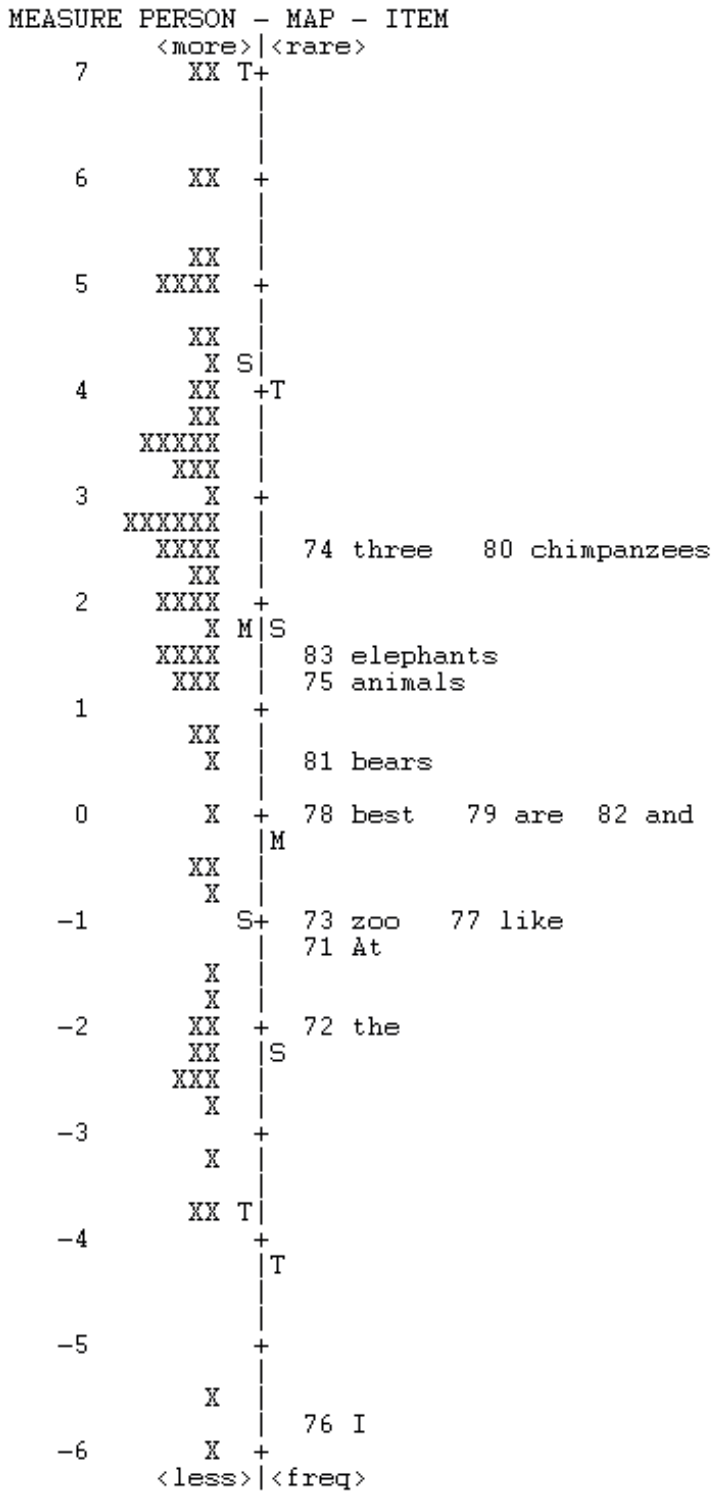


Figure 10 Item map for Form F

Form G

Table 20 Rasch item measures and corresponding Infit Mean Square statistics for Form G

Item	Measure (logits)	Infit MNSQ
84	-1.06	0.63
85	0.90	1.18
86	1.36	0.95
87	0.14	1.06
88	0.89	0.59
89	-0.09	0.77
90	-2.08	1.47
91	1.47	0.77
92	0.42	0.70
93	1.79	1.02
94	0.13	0.77
95	0.47	0.92
96	0.63	0.74
97	-0.22	1.05

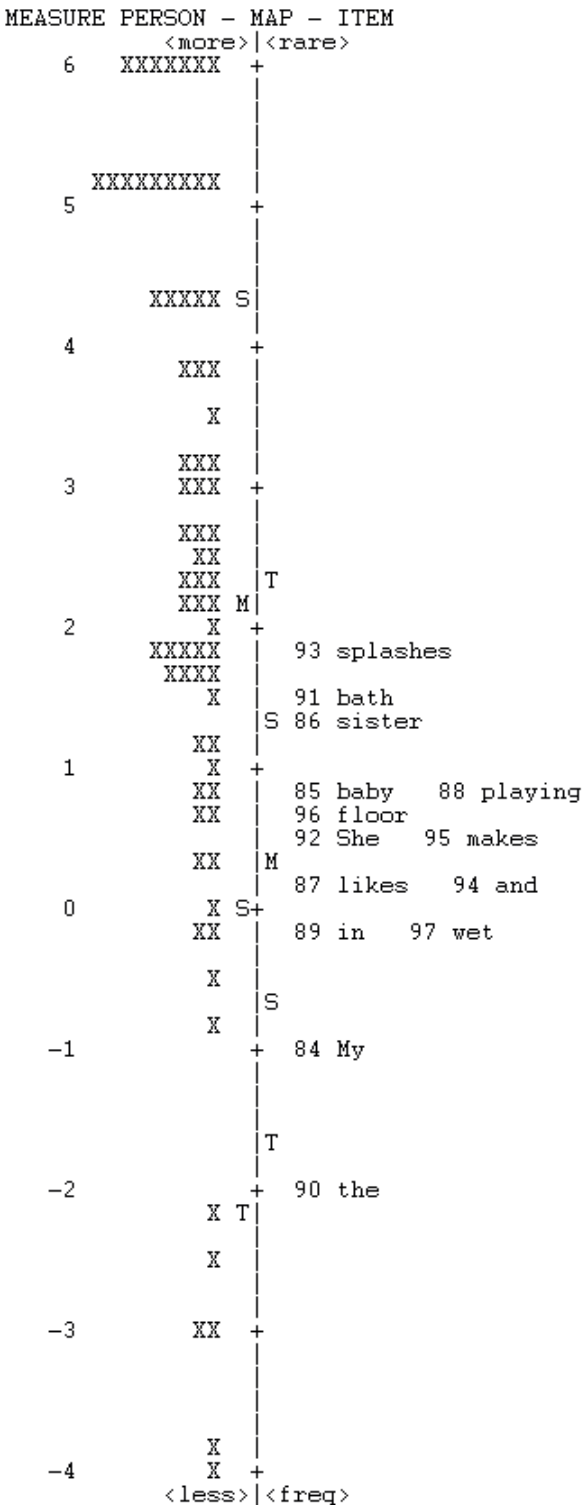
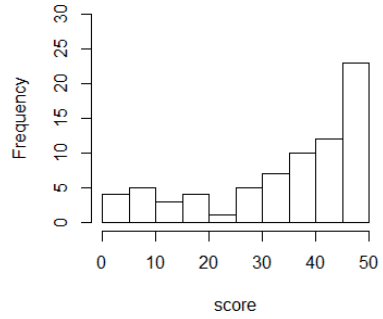


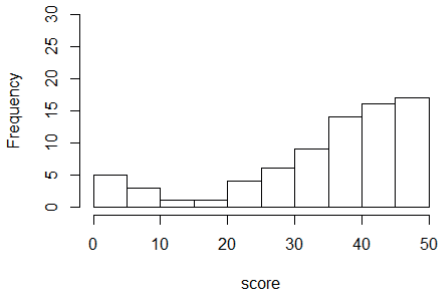
Figure 11 Item map for Form G

Appendix 2: Raw score distributions for the pilot forms

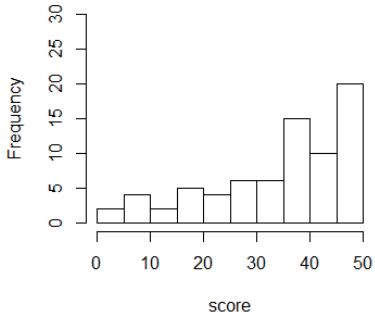
Form A



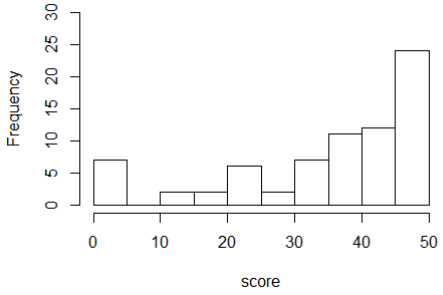
Form B



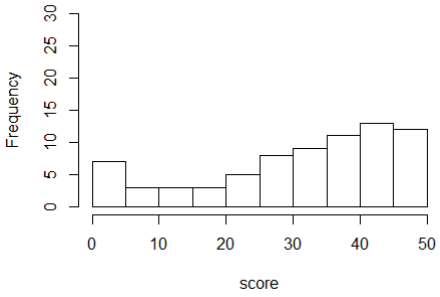
Form C



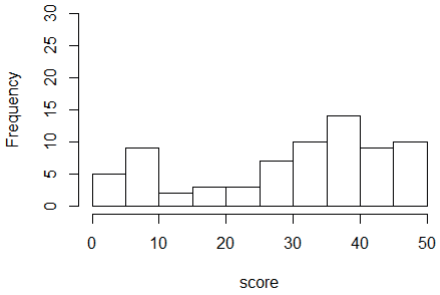
Form D



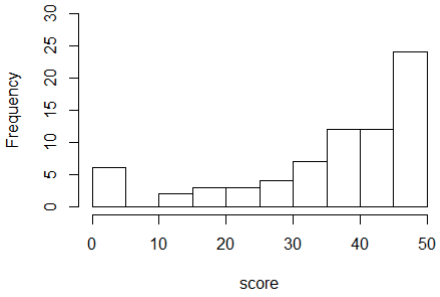
Form E



Form F



Form G



Appendix 3: Raw Score to stanine conversion tables for the current forms

Hearing and Recording Sounds in Words: 5:00-5:50 years

Test score	0	1	2-4	5-11	12-18	19-26	27-33	34-36	37
Stanine	1	2	3	4	5	6	7	8	9

Hearing and Recording Sounds in Words: 5:51-6:00 years

Test score	0-1	2-7	8-15	16-22	23-29	30-33	34-36	37	
Stanine	1	2	3	4	5	6	7	8	9

Hearing and Recording Sounds in Words: 6:01-6:50 years

Test score	0-8	9-19	20-27	28-32	33-35	36	37		
Stanine	1	2	3	4	5	6	7	8	9

Hearing and Recording Sounds in Words: 6:51-7:00 years

Test score	0-14	15-28	29-32	33-35	36	37			
Stanine	1	2	3	4	5	6	7	8	9

References

Clay, M.M. (2019). *An Observation Survey of Early Literacy Achievement* (4th ed.). Auckland, New Zealand: The Marie Clay Literacy Trust.

Clay, M. M. (2014). *By different paths to common outcomes*. Auckland, New Zealand: The Marie Clay Literacy Trust.

Foulin, J. N (2005). Why is letter-name knowledge such a good predictor of learning to read? *Reading and Writing*, 18, 129–155.

Goswami, U. (2008). Reading, complexity and the brain. *Literacy*, 42(2), 67–74.

Kim, Y.K., Petscher, Y., Foorman, B. R., & Zhou, C. (2010). The contributions of phonological awareness and letter-name knowledge to letter-sound acquisition: A cross-classified multilevel model approach. *Journal of Educational Psychology*, 102(2), 313–326.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.