Standardisation of the Observation Survey in England and Wales, UK

Andrew J. Holliman*, Jane Hurry**, & Julia Douetil**
Coventry University, UK* Institute of Education, University of London, UK**

Sampling procedure

Schools

It was important to ensure, as far as possible, that the sample of schools in this study accurately represented the wider population of schools in England and Wales. To maximise the chances of this, the National Foundation for Educational Research (NFER), which contains full details for all mainstream schools in England and Wales, were commissioned to generate a random stratified sample of three-hundred schools including the correct proportions of *government office region* (North-East, North-West, Yorkshire, East-Midlands, West-Midlands, Eastern, London, South-East, South-West, Wales), *percentage eligible for free school meals* (Lowest 20%, Second Lowest 20%, Middle 20%, Second Highest 20%, Highest 20%), and *school locale* (Urban, Rural).

Initial contact was made with each of these schools through letter correspondence. This was follwed by a telephone call and an email (if available) two weeks later seeking their participation. A systematic random procedure was followed for contacting schools; that is, every third school on the list of schools was contacted until all schools had been contacted at least once, to avoid any potential bias. Despite our best efforts, a total of sixty-nine schools were recruited for participation in this study. The sampling characteristics of this *reduced* sample represented the wider population of schools fairly well, although there was most notably an over-representation of schools from the South-East and London, along with an over-representation of Reading Recovery schools in this sample (see Table 1).

Testing was carried out towards the end of the 2008-2009 school year in June-July by Reading Recovery teachers, Reading Recovery Teacher Leaders, retired Reading Recovery teachers, newly qualified Reading Recovery teachers, and some other trained associates. The essential prerequisite for test administrators was that they should be trained at administering the Observation Survey and should have administered the Observation Survey within the last six months. Test administrators were instructed to follow administration instructions precisely to help ensure consistency. In most cases test administrators collected data from schools in their Local Authorities, although some travelled great distances to collect data in other schools. It was also not uncommon to have two-to-three Reading Recovery teachers working in a single school at any given time so that data could be collected in a single day (in most cases) and to minimise interference.

Children

In each of the sixty-nine participating schools, the aim was to select two boys and two girls from the following four age-ranges (in years):

- 5.00 - 5.50
- 5.51 - 6.00
- 6.01 - 6.50
- 6.51 - 7.00

These children were selected using a systematic random sampling procedure; that is, registers for Reception, Year 1, and Year 2 classes were arranged in alphabetical order by teachers' surname, and then every *ninth* child (chosen at random) was selected for possible participation. Once a child was identified, it was checked whether they were eligible for inclusion based on the criteria: two boys and two girls from the four age-ranges. This systematic random selection process was followed until sixteen appropriate children were identified from each school. Note: we decided not to include Reading Recovery children in this study due to their familiarity with the assessments. While this potentially introduced a selection bias, according to test administrators, the incidence rate of this was very low and therefore should not have affected the results. It should also be noted that for a very small sample of schools, this random selection process had to be abandoned due to too few participants matching our selection criteria.

It was equally important to ensure, as far as possible, that the sample of children tested at each of the four age-ranges accurately represented the wider population of children in England and Wales. The following information was obtained for each participating child: sex, ethnicity, first language, and whether the child is eligible for free school meals. The data from nine-hundred-and-eighty participating children were compared with the population data from the 2009 census by the Department for Children, Schools and Families (DCSF) found at http://www.dcsf.gov.uk/rsgateway/DB/SFR/s000843/index.shtml, using English schools only (see Table 2). The proportions of sex, ethnicity, first language, and eligibility for free school meals were all very well-represented in this sample.

The standardisation testing

Stanine scores

The following tasks of the Observation Survey were standardised: Letter Identification, Concepts About Print, Duncan Word Test, Writing Vocabulary, and Hearing and Recording Sounds in Words. Note: the Duncan Word Test was standardised because this is the preferred test used to assess reading in the Reading Recovery programme in the UK, as opposed to the Ohio Word Test or Clay Word Test. Also, while regrettable, it was also not possible given the resources available to standardise Text Reading Level in this study.

The stanines and percentile ranks were calculated for each test, for each of the four age-ranges (outlined previously). These results are displayed in the tables overleaf (see Tables 3-10). In these tables, 'N' refers to the number of participants (or observed scores), 'Mean' refers to the average of all the scores, 'SD' refers to standard deviation from the mean of all observed scores, and 'Range' shows the lowest and highest possible score.

Technical details

Although the definition of stanine is consistent in the literature in that the mean of the stanine scale is 5 with a standard deviation of 2, and that the percentage of cases for stanines 1-9 are 4, 7, 12, 17, 20, 17, 12, 7, and 4 (Crocker & Algina, 1986; McDaniel, 1994; Taylor, 2010), the actual cut-off corresponding percentile rank to stanines differ. The following standard provided by Crocker and Algina (1986) was used to put score values into each category of stanine according to the cumulative percent in the last column of the frequency distribution table.

Stanine	1	2	3	4	5	6	7	8	9
Percentile Rank (%)	< 4	4-10	11-22	23-39	40-59	60-76	77-88	89-95	>= 96
Percentage of Cases (%)	4	7	12	17	20	17	12	7	4

Since the population size was not big enough, there are some missing values for a particular category of stanine. In other cases, no observation score was observed as a continuum; for example, for the 6.51-7.00 age group, no one out of 223 students scored 21 for Hearing and Recording Sounds in Words (HRSW). This is not surprising. With the growth of population size, it is likely that at least one student would obtain each of the possible scores. For the convenience of teachers using the stanine table, missing values were added to the table even though they were not observed. The missing value was placed with the closest smaller available value to be conservative. The same procedures with missing values were followed for percentile ranks. The minimum and maximum scores were included even though they might not be in the data. Missing values in the middle were also included by treating them as the closest smaller value to be conservative.

Reliability and validity testing

Overview

It was important to investigate the reliability (consistency) and validity (whether the test measures what it purports to measure) of the Observation Survey in order to enhance its' credibility as a measure of early literacy. Therefore, in addition to calculating the stanines for each age-range, a series of analyses were performed to assess the various types of reliability and validity with this sample. To support this process, a subsample of approximately one-hundred-and-twelve participating children from eight schools, selected on a convenience basis, completed the Observation Survey on two separate occasions (no longer than one week apart) and also received an additional reading and spelling test on the second round of assessments.

The results from these reliability and validity analyses are displayed in the figures and tables on the subsequent pages, following the stanine tables. However, the major types of reliability and validity that were investigated in this study have been briefly introduced and discussed below.

Validity

Construct validity refers to whether a test measures the psychological characteristic of interest; in this case literacy. This has often been demonstrated by showing that children improve with age (as can be seen in Figure 1). Related to this, *content validity* refers to the extent to which the content of a test can be said to be representative of the skill or ability the test is designed to measure. If all tasks in the Observation Survey are tapping into the same skill, literacy, the component scales should correlate highly with the total score from all tests; this can be seen in Table 11. To investigate the relatedness of these tests even further, a factor analysis was conducted to investigate how they loaded together. The method used for factor extraction was 'Principle Components' and the rotation method was 'Varimax'. Using the 'Eigenvalues greater than 1' criterion along with observations of the Scree Plot, this analysis confirmed that the five measures in the Observation Survey loaded

heavily together (from .809 to .925) onto a single factor only, with the Duncan Word Test (.925) and the Hearing and Recording Sounds in Words Test (.895) most heavily loaded onto this factor. This suggests that the assessments in the Observation Survey are tapping into the same underlying construct.

Criterion validity refers to the relationship between a measure and other measures purporting to measure the same ability. If the Observation Survey measures literacy, we would expect to see strong correlations with other tests of literacy administered concurrently (at the same time). To investigate this, the Primary Reading Test (France, 1979) and the British Spelling Test Series (Vincent & Crumpler, 1997) were both administered to the subsample of children. These tests were selected for several reasons; they were appropriate for the targeted age-range, they have been previously standardised and assessed for their reliability and validity, and importantly for practical purposes, they could be group administered which had many advantages (e.g., enabling children to work at their own pace in a more relaxed atmosphere, and enabling data collectors to assess groups of children on a single occasion, thus minimising interference in schools). In the Primary Reading Test (questions 1-16) children were presented with a picture (e.g., of a bed) and were then asked by the test administrator, in accordance with the standardised instructions, to circle the word that went with that picture from a choice of five available (e.g., in this example, tap, bed, lid, tub, and pot). In the British Spelling Test Series, spelling was assessed at the word, sentence, and continuous writing level, and also involved other activities such as identifying spelling errors in words. As can be seen from Table 12, all assessments in the Observation Survey for all age-ranges were significantly correlated with spelling and reading (apart from a single case), although it should be noted that many correlations could not be described as 'strong' (i.e., they were <.7). This lack of strength could perhaps be explained in part by the nature of these tests; for instance, pictures are an integral part of the Primary Reading Test and the British Spelling Test Series and facilitate children's answers. While this format is especially useful for engaging beginning readers and for assessing literacy at its basic levels, it does introduce a potential bias by favouring those children with more developed receptive vocabularies. Thus, subtle differences in the domains being tested might, to some extent, have explained the predominantly 'moderate' correlations between tests. It is also regrettable that only the first 16 items were used to assess reading in the Primary Reading Test due to inconsistencies in administration, and this might also explain the moderate correlations in this study.

Reliability

In order to demonstrate the *test-retest (stability) reliability* of the Observation Survey, we would expect to find a strong correlation between scores on this test over two separate occasions. It is recommended that parallel forms of tests should be used to reduce practice effects and this was done where available (i.e., in the Concepts About Print Test, Duncan Word Test, and the Hearing and Recording Sounds in Words Test). As can be seen in Table 13, the correlations between performance over the two occasions were strong (>0.7) and significant, and this was the case for all assessments across all age-ranges.

'Internal reliability' refers to the extent to which the items within a test appear to be measuring the same skill, or underlying ability. As can be seen in Table 14, the internal reliability (using Cronbach's alpha) was very high (0.9 in most cases) and always above 0.6.

In summary, these additional analyses have shown the reliability and validity of the Observation Survey to be strong. This adds to its' credibility as a measure of early literacy.

Acknowledgements

We would like to thank all teacher leaders, Reading Recovery teachers, and all other research assistants for their support with the data collection process. We would also like to thank all schools, teachers, parents, and children for taking part in this research. We would like to pay a special thank you to Chuang Wang for developing the stanines and percentiles and for providing some technical information, along with Jeff Brymer-Bashore for his support throughout this process.

References

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth Group/Thomson Learning.

McDaniel, E. (1994). *Understanding educational measurement.* New York: McGraw Hill.

Taylor, S. (2010). *Explanation of our now I understand approach to standardized testing: TerraNova and its scoring.* Retreived May 27, 2010 from http://www.itesteval.com/resources.shtml

Table 1: Sample characteristics of schools: Obtained percentages compared with the NFER (2009) register

	Population	Sample
Geographical region		
Wales	6.4	1.4
North-East	5.2	2.9
North-West	14.5	5.8
Yorkshire and the Humber	10.0	13.0
East-Midlands	8.7	4.3
West-Midlands	10.2	8.7
Eastern	10.7	7.2
London	10.2	18.8
South-East	14.0	24.6
South-West	10.1	13.0
Percentage eligible for FSM		
Lowest 20%	19.8	14.5
2nd lowest 20%	19.7	23.2
Middle 20%	18.3	20.3
2nd highest 20%	18.1	23.2
Highest 20%	17.8	17.4
Urban/Rural		
Rural	19.6	9.6
Non-Rural	80.3	63.8
Type of school		
RR school	7.1	20.3
Non RR school	92.9	79.7
Note: The above percentages are based on the data we have and do not include missing data.		
The 'Type of school' data were calculated by hand using information from the NFER register, along with information from the Reading Recovery National Network (2008-2009).		

Table 2: Sample characteristics of pupils: Obtained percentages compared with the DCSF (2009) census

	Population	Sample
Sex*		
Males	51.1	49.3
Females	48.9	50.7
Ethnicity**		
White: Eastern European	Not specified	4.6
Any other white background	Not specified	73.4
Mixed: White and black Caribbean	1.3	1.6
Mixed: White and black African	0.5	0.3
Mixed: White & Asian	0.9	1.1
Mixed: Any other mixed background	1.5	1.0
Asian: Indian	2.5	3.2
Asian: Pakistani	3.9	3.4
Asian: Bangladeshi	1.6	1.4
Asian: Any other Asian background	1.3	1.5
Black: Caribbean	1.4	1.1
Black: African	2.9	3.5
Black: Any other black background	0.6	0.9
Chinese: Chinese	0.3	0.6
Chinese: Japanese	Not specified	0.0
Other	1.4	1.5
Unknown	NA	0.8
First language**		
English	84.6	84.8
Other	15.2	15.2
Unclassified	0.2	NA
Is child eligible for FSM**		
Yes	16.0	15.9
No	84.0	84.1
Note: The above percentages were obtained from the DCSF (2009) census and are based either on		
5-7 years olds only* or primary school children collectively**.		
The percentages are based on the data we have and do not include missing data unless specified.		

Table 3: Stanine scores for 5.00 - 5.50 years (for 254 children in 2009)

Letter Identification (LI)									
Purpose: To find what letters a child knows and the preferred mode of identification.									
Task: Identify upper- and lower-case letters and print forms of "a" and "g".									
Scoring: *N* = 254; *Mean* = 45.01; *SD* = 9.28; *Range* = 0-54.									
Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-23	24-30	31-41	42-46	47-49	50	51-52	~	53-54

Concept About Print (CAP)									
Purpose: To find what a child has learned about how spoken language is put into print.									
Task: Perform a variety of tasks during book reading by the teacher.									
Scoring: *N* = 254; *Mean* = 12.35; *SD* = 3.74; *Range* = 0-24.									
Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-5	6-7	8	9-11	12	13-14	15-16	17	18-24

Duncan Word Test (DWT)									
Purpose: To find if a child is developing a personal resource of reading vocabulary.									
Task: Read a list of high-frequency words.									
Scoring: *N* = 254; *Mean* = 9.93; *SD* = 6.58; *Range* = 0-23.									
Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	~	0-1	2-3	4-7	8-11	12-14	15-19	20-21	22-23

Writing Vocabulary (WV)									
Purpose: To find if a child is building a personal resource of words that can be written.									
Task: Write all known words in 10 minutes.									
Scoring: *N* = 250; *Mean* = 12.68; *SD* = 10.08; *Range* = 0-MAX.									
Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	~	0-1	2-3	4-7	8-12	13-18	19-24	25-30	31+

Hearing and Recording Sounds in Words (HRSIW)									
Purpose: To assess phonemic awareness by determining how well a child represents the sounds of letters									
and clusters of letters in graphic form.									
Task: Write a dictated sentence, with credit for sounds correctly represented.									
Scoring: *N* = 250; *Mean* = 22.28; *SD* = 10.77; *Range* = 0-37.									
Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-1	2-5	6-11	12-20	21-27	28-31	32-34	35	36-37

Table 4: Stanine and percentile ranks for 5.00 - 5.50 years (for 254 children in 2009)

Letter Identification (*N* = 254)			Concept About Print (N = 254)			Duncan Word test (N = 254)			Writing Vacabulary (N = 254)			HRSW (N = 254)		
Raw	Stan.	Perc.	Raw	Stan.	Perc.	Raw	Stan.	Perc.	Raw	Stan.	Perc.	Raw	Stan.	Perc.
0-11	1	0	0-2	1	0	0	2	6	0	2	4	0-1	1	2
12-15	1	1	3	1	2	1	2	10	1	2	8	2	2	4
16-18	1	2	4-5	1	3	2	3	15	2	3	12	3	2	6
19-23	1	3	6	2	5	3	3	19	3	3	18	4	2	8
24-25	2	4	7	2	10	4	4	25	4	4	23	5	2	9
26-27	2	5	8	3	14	5	4	31	5	4	28	6	3	11
28-29	2	7	9	4	23	6	4	35	6	4	33	7	3	12
30	2	10	10	4	26	7	4	39	7	4	36	8	3	14
31-32	3	11	11	4	37	8	5	44	8	5	40	9	3	17
33	3	12	12	5	49	9	5	49	9	5	44	10	3	18
34	3	13	13	6	61	10	5	55	10	5	48	11	3	21
35	3	14	14	6	71	11	5	59	11	5	52	12	4	24
36	3	15	15	7	79	12	6	66	12	5	57	13	4	25
37	3	17	16	7	87	13	6	70	13	6	60	14	4	28
38	3	18	17	8	92	14	6	76	14	6	66	15	4	29
39	3	20	18	9	96	15	7	79	15	6	68	16	4	32
40	3	21	19-20	9	98	16	7	81	16	6	71	17	4	33
41	3	22	21-24	9	99	17	7	83	17	6	74	18	4	34
42	4	24				18	7	85	18	6	76	19	4	36
43	4	26				19	7	88	19	7	78	20	4	39
44	4	29				20	8	90	20	7	80	21	5	40
45	4	33				21	8	93	21	7	83	22	5	42
46	4	38				22	9	97	22	7	84	23	5	46
47	5	43				23	9	99	23	7	87	24	5	49
48	5	52							24	7	88	25	5	52
49	5	59							25	8	89	26	5	55
50	6	68							26	8	90	27	5	58
51	7	77							27	8	91	28	6	63
52	7	88							28	8	92	29	6	66
53	9	97							29-30	8	94	30	6	70
54	9	99							31-33	9	96	31	6	74
									34-37	9	97	32	7	78
									38-44	9	98	33	7	83
									45+	9	99	34	7	88
												35	8	92
												36	9	97
												37	9	99

Note: Raw = Raw Score, Stan. = Stanine, Perc. = Percentile Rank; HRSW = Hearing and recording Sounds in Words (also known as Dictation).

Table 5: Stanine scores for 5.51 - 6.00 years (for 251 children in 2009)

Letter Identification (LI)									
Purpose: To find what letters a child knows and the preferred mode of identification.									
Task: Identify upper- and lower-case letters and print forms of "a" and "g".									
Scoring: *N* = 251; *Mean* = 48.47; *SD* = 8.15; *Range* = 0-54.									
Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-24	25-41	42-46	47-49	50-51	52	53	54	54
Concept About Print (CAP)									
Purpose: To find what a child has learned about how spoken language is put into print.									
Task: Perform a variety of tasks during book reading by the teacher.									
Scoring: *N* = 251; *Mean* = 15.00; *SD* = 4.31; *Range* = 0-24.									
Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-6	7-8	9-11	12-13	14-16	17	18-19	20	21-24
Duncan Word Test (DWT)									
Purpose: To find if a child is developing a personal resource of reading vocabulary.									
Task: Read a list of high-frequency words.									
Scoring: *N* = 249; *Mean* = 15.41; *SD* = 7.37; *Range* = 0-23.									
Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	~	0-3	4-8	9-14	15-19	20-21	22	23	23
Writing Vocabulary (WV)									
Purpose: To find if a child is building a personal resource of words that can be written.									
Task: Write all known words in 10 minutes.									
Scoring: *N* = 249; *Mean* = 22.22; *SD* = 14.70; *Range* = 0-MAX.									
Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	~	0-4	5-9	10-16	17-23	24-32	33-39	40-51	52+
Hearing and Recording Sounds in Words (HRSIW)									
Purpose: To assess phonemic awareness by determining how well a child represents the sounds of letters									
and clusters of letters in graphic form.									
Task: Write a dictated sentence, with credit for sounds correctly represented.									
Scoring: *N* = 248; *Mean* = 28.14; *SD* = 9.94; *Range* = 0-37.									
Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-2	3-10	11-22	23-29	30-33	34-35	36	37	37

Table 6: Stanine and percentile ranks for 5.51 - 6.00 years (for 251 children in 2009)

Letter Identification (*N* = 251)			Concept About Print (N = 251)			Duncan Word test (N = 249)			Writing Vacabulary (N = 249)			HRSW (N = 248)		
Raw	Stan.	Perc.	Raw	Stan.	Perc.	Raw	Stan.	Perc.	Raw	Stan.	Perc.	Raw	Stan.	Perc.
0-9	1	0	0-3	1	1	0	2	4	0	2	4	0	1	1
10-19	1	1	4-5	1	2	1	2	7	1	2	6	1	1	2
20-23	1	2	6	1	3	2	2	9	2-3	2	8	2	1	3
24	1	3	7	2	6	3	2	10	4	2	9	3-6	2	4
25-26	2	4	8	2	8	4	3	12	5	3	11	7	2	5
27-30	2	5	9	3	12	5	3	14	6	3	12	8-9	2	8
31-34	2	6	10	3	14	6	3	16	7	3	14	10	2	9
35-37	2	7	11	3	18	7	3	19	8	3	18	11	3	11
38-40	2	8	12	4	25	8	3	20	9	3	20	12-14	3	12
41	2	10	13	4	30	9	4	23	10	4	23	15	3	13
42	3	11	14	5	40	10	4	25	11	4	24	16	3	14
43	3	13	15	5	51	11	4	29	12	4	27	17	3	15
44	3	16	16	5	59	12	4	32	13	4	29	18	3	17
45	3	17	17	6	71	13	4	35	14	4	33	19	3	19
46	3	19	18	7	77	14	4	37	15	4	34	20-21	3	20
47	4	23	19	7	88	15	5	41	16	4	39	22	3	21
48	4	26	20	8	91	16	5	43	17	5	42	23	4	23
49	4	29	21	9	96	17	5	47	18	5	47	24	4	26
50	5	40	22-23	9	98	18	5	53	19	5	51	25	4	27
51	5	55	24	9	99	19	5	56	20	5	52	26	4	29
52	6	70				20	6	62	21	5	55	27	4	31
53	7	86				21	6	71	22	5	56	28	4	34
54	8-9	99				22	7	83	23	5	59	29	4	36
						23	8-9	99	24	6	61	30	5	42
									25	6	63	31	5	46
									26	6	65	32	5	52
									27	6	66	33	5	57
									28	6	67	34	6	67
									29	6	69	35	6	76
									30	6	71	36	7	85
									31	6	74	37	8-9	99
									32	6	76			
									33	7	77			
									34	7	79			
									35	7	82			
									36	7	84			
									37	7	86			
									38	7	87			
									39	7	88			
									40	8	89			
									41-42	8	90			
									43	8	93			
									44-46	8	94			
									47-51	8	95			
									52-55	9	96			
									56	9	97			
									57-67	9	98			
									68+	9	99			

Note: Raw = Raw Score, Stan. = Stanine, Perc. = Percentile Rank; HRSW = Hearing and recording Sounds in Words (also known as Dictation).

Table 7: Stanine scores for 6.01 - 6.50 years (for 248 children in 2009)

Letter Identification (LI)									
Purpose: To find what letters a child knows and the preferred mode of identification.									
Task: Identify upper- and lower-case letters and print forms of "a" and "g".									
Scoring: *N* = 248; *Mean* = 51.26; *SD* = 3.63; *Range* = 0-54.									
Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-44	45-47	48-49	50-51	52	~	53	54	54

Concept About Print (CAP)									
Purpose: To find what a child has learned about how spoken language is put into print.									
Task: Perform a variety of tasks during book reading by the teacher.									
Scoring: *N* = 249; *Mean* = 17.48; *SD* = 3.56; *Range* = 0-24.									
Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-9	10-12	13-14	15-16	17	18-20	21	22	23-24

Duncan Word Test (DWT)									
Purpose: To find if a child is developing a personal resource of reading vocabulary.									
Task: Read a list of high-frequency words.									
Scoring: *N* = 247; *Mean* = 19.58; *SD* = 4.70; *Range* = 0-23.									
Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-8	9-12	13-16	17-20	21	22	23	23	23

Writing Vocabulary (WV)									
Purpose: To find if a child is building a personal resource of words that can be written.									
Task: Write all known words in 10 minutes.									
Scoring: *N* = 244; *Mean* = 29.97; *SD* = 15.06; *Range* = 0-MAX.									
Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-8	9-13	14-18	19-23	24-31	32-39	40-48	49-59	60+

Hearing and Recording Sounds in Words (HRSIW)									
Purpose: To assess phonemic awareness by determining how well a child represents the sounds of letters									
and clusters of letters in graphic form.									
Task: Write a dictated sentence, with credit for sounds correctly represented.									
Scoring: *N* = 244; *Mean* = 33.95; *SD* = 4.11; *Range* = 0-37.									
Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-23	24-28	29-31	32-34	35	36	37	37	37

Table 8: Stanine and percentile ranks for 6.01 - 6.50 years (for 248 children in 2009)

Letter Identification (*N* = 248)			Concept About Print (N = 249)			Duncan Word test (N = 247)			Writing Vacabulary (N = 244)			HRSW (N = 244)		
Raw	Stan.	Perc.	Raw	Stan.	Perc.	Raw	Stan.	Perc.	Raw	Stan.	Perc.	Raw	Stan.	Perc.
0-39	1	0	0-8	1	0	0-3	1	0	0-3	1	0	0-17	1	0
40	1	1	9	1	2	4-5	1	1	4-6	1	1	18	1	1
41-43	1	2	10	2	4	6-7	1	2	7	1	2	19-21	1	2
44	1	3	11	2	5	8	1	3	8	1	3	22-23	1	3
45	2	4	12	2	9	9	2	5	9	2	4	24-25	2	4
46	2	6	13	3	14	10	2	6	10-11	2	7	26	2	6
47	2	9	14	3	18	11	2	8	12	2	8	27-28	2	7
48	3	12	15	4	26	12	2	10	13	2	10	29	3	11
49	3	19	16	4	36	13	3	13	14	3	13	30	3	15
50	4	27	17	5	48	14	3	16	15	3	14	31	3	19
51	4	34	18	6	62	15	3	19	16	3	17	32	4	24
52	5	57	19	6	70	16	3	21	17	3	19	33	4	30
53	7	77	20	6	76	17	4	25	18	3	22	34	4	37
54	8-9	99	21	7	85	18	4	29	19	4	24	35	5	50
			22	8	92	19	4	32	20	4	29	36	6	68
			23	9	98	20	4	37	21	4	32	37	7-9	99
			24	9	99	21	5	43	22	4	35			
						22	6	60	23	4	38			
						23	7-9	99	24	5	43			
									25	5	45			
									26	5	50			
									27-28	5	52			
									29	5	54			
									30	5	55			
									31	5	58			
									32	6	61			
									33	6	63			
									34	6	67			
									35	6	69			
									36	6	72			
									37	6	73			
									38	6	74			
									39	6	76			
									40	7	78			
									41	7	79			
									42	7	82			
									43	7	83			
									44	7	84			
									45-46	7	87			
									47-48	7	88			
									49	8	89			
									50	8	90			
									51-54	8	91			
									55	8	92			
									56-57	8	93			
									58	8	94			
									59	8	95			
									60	9	96			
									61-62	9	97			
									63-75	9	98			
									76+	9	99			

Note: Raw = Raw Score, Stan. = Stanine, Perc. = Percentile Rank; HRSW = Hearing and recording Sounds in Words (also known as Dictation).

Table 9: Stanine scores for 6.51 - 7.00 years (for 224 children in 2009)

Letter Identification (LI)

Purpose: To find what letters a child knows and the preferred mode of identification.

Task: Identify upper- and lower-case letters and print forms of "a" and "g".

Scoring: *N* = 224; *Mean* = 51.85; *SD* = 3.31; *Range* = 0-54.

Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-44	45-48	49-50	51-52	~	53	54	54	54

Concept About Print (CAP)

Purpose: To find what a child has learned about how spoken language is put into print.

Task: Perform a variety of tasks during book reading by the teacher.

Scoring: *N* = 224; *Mean* = 18.92; *SD* = 3.39; *Range* = 0-24.

Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-11	12-14	15-16	17	18-19	20-21	22	23	24

Duncan Word Test (DWT)

Purpose: To find if a child is developing a personal resource of reading vocabulary.

Task: Read a list of high-frequency words.

Scoring: *N* = 224; *Mean* = *20.80*; *SD* = 4.18; *Range* = 0-23.

Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-8	9-15	16-19	20-21	22	23	23	23	23

Writing Vocabulary (WV)

Purpose: To find if a child is building a personal resource of words that can be written.

Task: Write all known words in 10 minutes.

Scoring: *N* = 224; *Mean* = 36.56; *SD* = 16.59; *Range* = 0-MAX.

Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-8	9-16	17-22	23-29	30-39	40-48	49-58	59-67	68+

Hearing and Recording Sounds in Words (HRSIW)

Purpose: To assess phonemic awareness by determining how well a child represents the sounds of letters and clusters of letters in graphic form.

Task: Write a dictated sentence, with credit for sounds correctly represented.

Scoring: *N* = 223; *Mean* = 34.49; *SD* = 5.18; *Range* = 0-37.

Stanine Group	1	2	3	4	5	6	7	8	9
Test Score	0-21	22-30	31-33	34-35	36	37	37	37	37

Table 10: Stanine and percentile ranks for 6.51 - 7.00 years (for 224 children in 2009)

Letter Identification (N = 224)			Concept About Print (N = 224)			Duncan Word test (N = 224)			Writing Vacabulary (N = 224)			HRSW (N = 223)		
Raw	Stan.	Perc.	Raw	Stan.	Perc.	Raw	Stan.	Perc.	Raw	Stan.	Perc.	Raw	Stan.	Perc.
0-35	1	0	0-8	1	0	0-2	1	0	0-5	1	0	0-8	1	0
36-41	1	1	9-10	1	1	3-6	1	1	6-7	1	2	9-12	1	1
42	1	2	11	1	2	7	1	2	8	1	3	13-19	1	2
43-44	1	3	12	2	4	8	1	3	9-10	2	4	20-21	1	3
45	2	4	13	2	6	9-10	2	4	11-12	2	5	22-23	2	4
46	2	5	14	2	8	11	2	6	13	2	6	24	2	5
47	2	6	15	3	14	12-13	2	7	14	2	7	25-26	2	6
48	2	8	16	3	20	14	2	8	15	2	8	27-28	2	7
49	3	12	17	4	30	15	2	10	16	2	10	29	2	8
50	3	18	18	5	41	16	3	11	17-18	3	12	30	2	9
51	4	28	19	5	52	17	3	12	19	3	15	31	3	13
52	4	39	20	6	63	18	3	15	20	3	16	32	3	16
53	6	69	21	6	75	19	3	18	21	3	18	33	3	19
54	7-9	99	22	7	87	20	4	23	22	3	21	34	4	25
			23	8	92	21	4	31	23-24	4	25	35	4	34
			24	9	99	22	5	45	25	4	28	36	5	49
						23	6-9	99	26	4	31	37	6-9	99
									27	4	33			
									28	4	34			
									29	4	36			
									30	5	41			
									31-32	5	42			
									33	5	44			
									34	5	46			
									35	5	47			
									36	5	51			
									37	5	54			
									38	5	56			
									39	5	58			
									40	6	62			
									41	6	64			
									42	6	66			
									43	6	67			
									44	6	70			
									45	6	71			
									46	6	72			
									47	6	73			
									48	6	75			
									49	7	78			
									50	7	79			
									51	7	82			
									52	7	83			
									53	7	84			
									54	7	86			
									55-57	7	87			
									58	7	88			
									59	8	89			
									60-61	8	90			
									62	8	92			
									63	8	93			
									64	8	94			
									65-67	8	95			
									68-69	9	96			
									70-71	9	97			
									72-76	9	98			
									77+	9	99			

Note: Raw = Raw Score, Stan. = Stanine, Perc. = Percentile Rank; HRSW = Hearing and recording Sounds in Words (also known as Dictation).

Figure 1: Score distributions by age group

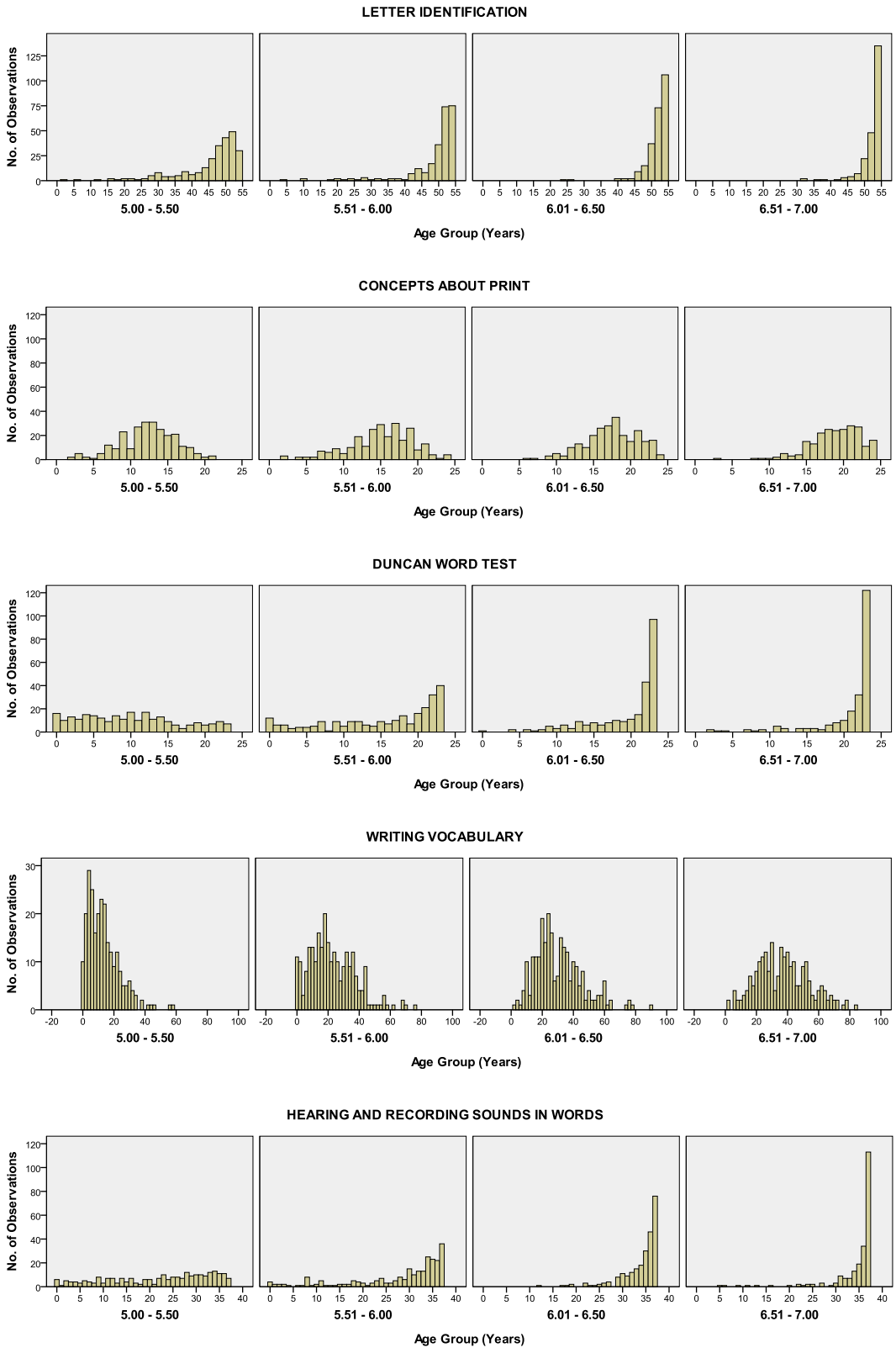

Table 11: Inter-correlations among tasks by age group

Inter-correlations among Observation Survey scores for 5.00 - 5.50 years (n = 254)					
	Letter identification	Concepts about print	Duncan word test	Writing vocabulary	HRSW
Letter identification	1.000	0.484***	0.655***	0.514***	0.662***
Concepts about print	0.484***	1.000	0.612***	0.444***	0.584***
Duncan word test	0.655***	0.612***	1.000	0.660***	0.721***
Writing vocabulary	0.514***	0.444***	0.660***	1.000	0.636***
HRSW	0.662***	0.584***	0.721***	0.636***	1.000
Inter-correlations among Observation Survey scores for 5.51 - 6.00 years (n = 251)					
	Letter identification	Concepts about print	Duncan word test	Writing vocabulary	HRSW
Letter identification	1.000	0.653***	0.682***	0.504***	0.712***
Concepts about print	0.653***	1.000	0.758***	0.596***	0.667***
Duncan word test	0.682***	0.758***	1.000	0.677***	0.757***
Writing vocabulary	0.504***	0.596***	0.677***	1.000	0.623***
HRSW	0.712***	0.667***	0.757***	0.623***	1.000
Inter-correlations among Observation Survey scores for 6.01 - 6.50 years (n = 248)					
	Letter identification	Concepts about print	Duncan word test	Writing vocabulary	HRSW
Letter identification	1.000	0.397***	0.496***	0.327***	0.460
Concepts about print	0.397***	1.000	0.669***	0.562***	0.527***
Duncan word test	0.496***	0.669***	1.000	0.509***	0.710***
Writing vocabulary	0.327***	0.562***	0.509***	1.000	0.510***
HRSW	0.460***	0.527***	0.710***	0.510***	1.000
Inter-correlations among Observation Survey scores for 6.50 - 7.00 years (n = 224)					
	Letter identification	Concepts about print	Duncan word test	Writing vocabulary	HRSW
Letter identification	1.000	0.482***	0.638***	0.376***	0.641***
Concepts about print	0.482***	1.000	0.649***	0.621***	0.544***
Duncan word test	0.638***	0.649***	1.000	0.531***	0.725***
Writing vocabulary	0.376***	0.621***	0.531***	1.000	0.484***
HRSW	0.641***	0.544***	0.725***	0.484***	1.000
Inter-correlations among Observation Survey scores for the total sample (n = 967)					
	Letter identification	Concepts about print	Duncan word test	Writing vocabulary	HRSW
Letter identification	1.000	0.594***	0.694***	0.501***	0.720***
Concepts about print	0.594***	1.000	0.777***	0.682***	0.692***
Duncan word test	0.694***	0.777***	1.000	0.697***	0.804***
Writing vocabulary	0.501***	0.682***	0.697***	1.000	0.638***
HRSW	0.720***	0.692***	0.804***	0.638***	1.000
Note: For all tables, *p<.05, **p<.01, ***p<.001.					
Not all participants within each group completed all assessments.					

Table 12: Correlation between Observation Survey scores and measures of reading and spelling

	N	Primary Reading Test /16	N	British Spelling Test Series
5.00 - 5.50				
Letter identification	33	.699***	34	.527**
Concepts about print	33	.465**	34	.591***
Duncan word test	33	.705***	34	.797***
Writing vocabulary	33	.678***	34	.863***
HRSW	33	.702***	34	.695***
5.51 - 6.00				
Letter identification	34	.712***	36	.512**
Concepts about print	34	.747***	36	.694***
Duncan word test	34	.680***	36	.789***
Writing vocabulary	33	.496**	35	.798***
HRSW	34	.540**	36	.655***
6.01 - 6.50				
Letter identification	32	.380*	31	.564**
Concepts about print	32	.391*	31	.431*
Duncan word test	32	.515**	31	.697***
Writing vocabulary	32	.406*	31	.705***
HRSW	32	.467**	31	.571**
6.51 - 7.00				
Letter identification	23	.445*	24	.419*
Concepts about print	23	.513**	24	.697***
Duncan word test	23	.462*	24	.825***
Writing vocabulary	23	0.336	24	.567**
HRSW	23	.455*	24	.826***
All children				
Letter identification	125	.737***	125	.524***
Concepts about print	125	.658***	125	.724***
Duncan word test	125	.731***	125	.802***
Writing vocabulary	124	.579***	124	.799***
HRSW	125	.730***	125	.663***
Notes: *p<.05, **p<.01, ***p<.001.				
HRSW = Hearing and Recording Sounds in Words (also known as Dictation).				
Due to inconsistencies in administration the Primary Reading Test was scored using the first 16 questions only, rather than the whole test.				

Table 13: Test-retest reliability coefficients for each test in the Observation Survey by age group

Task	N	Letter identification	Concepts about print	Duncan word test	Writing vocabulary	HRSW
5.00 - 5.50	34	.938***	.864***	.920***	.938***	.944***
5.51 - 6.00	35	.974***	.863***	.951***	.882***	.866***
6.01 - 6.50	33	.743***	.854***	.921***	.819***	.776***
6.51 - 7.00	26	.715***	.881***	.936***	.914***	.939***
All children	128	.956***	.907***	.956***	.915***	.934***
Notes: *p<.05, **p<.01, ***p<.001.						
Alternate forms were used wherever possible (i.e. for the Concepts about print task, the Duncan word test, and the HRSW task).						
HRSW = Hearing and recording sounds in words (also known as dictation).						

Table 14: Internal consistency coefficients for the Observation Survey by age group

Task	N	Letter identification	Concepts about print	Duncan word test	HRSW
5.00 - 5.50	20	0.973	0.841	0.961	0.966
5.51 - 6.00	20	0.794	0.815	0.962	0.921
6.01 - 6.50	35	0.772	0.734	0.919	0.906
6.51 - 7.00	37	0.667	0.676	0.9	0.838
All children	112	0.935	0.832	0.955	0.952
Notes: Cronbach's alpha reliability coefficient was used to assess internal consistency.					
It was not possible to assess internal consistency for the Writing Vocabulary task.					
HRSW = Hearing and recording sounds in words (also known as dictation).					