NISS

Offline Reading Baseline Assessment: Combining 4th and 8th Grade NAEP Items

Analysis of ORCA 1 Study Data

Weiwei Cui, Eli Bruner, Nell Sedransk

Technical Report 190
May 2013

National Institute of Statistical Sciences
19 T.W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709
www.niss.org

Technical Report

Offline Reading Baseline Assessment: Combining 4th and 8th Grade NAEP Items

Analysis of ORCA I Study Data

W. Cui, E. Bruner, N. Sedransk
National Institute of Statistical Sciences

May 2013

ORCA Project

Don Leu, Project Leader
Jonna M. Kulikowich, Julio Coiro, Nell Sedransk Co-Principal Investigators
University of Connecticut ORCA Team Members

Offline Reading Baseline Assessment

Introduction

Internet learning, particularly the assessment of online learning skills, is receiving increasing attention throughout the education community at every level from grade-school classrooms to national forums and universities. Online reading comprehension can be thought of as a process involving several aspects of cognition in sequence, quite probably differentiating itself from the processes required for reading comprehension of printed text. That question can be thought of in terms of whether the online reading comprehension process of is driven by singular or multiple factors. Offline performance-based testing is a viable measure for the comprehension processes, including major sub-processes. Having created and implemented the Online Reading Comprehension Assessment (ORCA) to 1386 students in two states, each process can be examined on its own and in relation to other processes. As such, the purpose of this examination is to look at NAEP-based assessment in context with the ORCA Project is to develop a suitable standard offline scientific-reading testing standards. This is examined across public and private schools, as well as across states.

Method and Procedure

The ORCA Project test itself was created in four versions, based on selections from either of two 4^{th}-grade NAEP passages ("Wombats" and "Blue Crabs") and either of two 8^{th}-grade passages ("Cheater Meters" and "Sharebots"). Each of the four pairings was administered in both orders (i.e. a 4^{th}-grade passage followed by an 8^{th}-grade passage, or vise versa), totaling eight different Booklets (Table 1). Each assessment itself had seven items, each scored dichotomously – 4^{th}-grade passages had seven possible points to garner; 8^{th}-grade passages were functionally scored out of nine points. It is important to note, though, that the 8^{th}-grade Sharebots passage was initially to be scored out of 11. Questions three and six were originally scored 0-2; but their scoring was reduced to 0-1 after full credit for those items were not achieved by any of the students.

Table 1: Test design of the offline reading assessment

Booklet	Passage	Number of items	Total number of score points
1	Wombats+Cheater Meters	15	16
2	Cheater Meters+Wombats	15	16
3	Blue Crabs+Cheater Meters	15	16
4	Cheater Meters + Blue Crabs	15	16
5	Wombats+Sharebots	14	16*
6	Sharebots+Wombats	14	16*
7	Blue Crabs +Sharebots	14	16*
8	Sharebots + Blue Crabs	14	16*

*NAEP rubric for scoring for Sharebots would indicate 18 possible points for these pairs

Each examinee was given one of the 8 booklets. Responses were coded as 0 or 1, or 0, 1, or 2 where appropriate (i.e., some responses were coded as partially correct, and were given 1 point, while the fully correct answers were coded as 2). Reading scores on each passage were calculated as a sum of the score points on all items given to each examinee. Statistical and psychometric analyses were then conducted on both item and test levels

Statistical and Psychometric Analyses of Offline Reading Assessment

For Blue Crabs and Wombats passages, all items were dichotomously scored (scored 0 or 1). For Cheater Meter passage, 7 items were dichotomously scored (scored 0 or 1), one item was polytomously scored (i.e., scored 0, 1, or 2) using the NAEP scoring rubric. For Sharebots passage, originally (following NAEP) 3 items were dichotomously scored, while 4 items were polytomously scored. Later polytomous scoring of 2 items was reduced to dichotomous. Total score for each passage was calculated as a sum of the score points on each item.

1. Test difficulty

The overall percentages of correct responses to reading items are given in Table 1.1. As responses to the offline reading assessment were collected from 7th grade students, the 8th grade passage of Sharebots covers larger range of the reading domain compared with the other 8th grade passage of Cheater Meters (see Table 1.1). For the 4th grade passages, the passage of Blue Crabs has a more clustered difficulty distribution.

Table 1.1: Percent correct on reading items in offline reading assessment

Grade 4				Grade 8			
Wombats item	Percent correct	Blue Crabs item	Percent correct	Cheater Meters item	Percent correct	Sharebots item	Percent correct
WB1S	49.9	BC1S	86.1	CM1S (score =1)	47.0	SB1S	80.3
WB2S	97.1	BC2S	71.6	CM1S (score=2)	36.7	SB2S(score =1)	50.8
WB3S	69.8	BC3S	79.6	CM2S	87.0	SB2S(score=2)	16.2
WB4S	81.4	BC4S	74.3	CM3S	80.7	SB3S(score =1)	69.1
WB5S	89.1	BC5S	89.3	CM4S	55.9	SB3S(score=2)*	0
WB6S	78.0	BC6S	89.5	CM5S	84.4	SB4S(score =1)	55.4
WB7S	83.8	BC7S	89.7	CM6S	89.5	SB4S(score=2)	9.9
				CM7S	74.5	SB5S	52.8
				CM8S	73.6	SB6S(score =1)	69.4
						SB6S(score=2)*	0
						SB7S	16.4

*Revised scoring as 0 or 1

Score distributions for all four passages are shown in Table 1.2. All four passages have floor effects, and three of them show ceiling effects. Only the 8th grade passage of Sharebots has no

ceiling effect. As both 8th grade passages have floor effects, 4th grade reading materials must be included in the offline reading assessment in order to have an adequate measure of 7th grade students reading comprehension, in particular to distinguish among students with performance below the median. Within the two 8th grade passages, Sharebots has a better score distribution than the other 8th grade passage of Cheater Meters, which does not exhibit a ceiling effect.

Table 1.2: Table of distribution of total scores by order of presentation (**Frequency** and Percentage)

SCORE	Grade 4				Grade 8			
	Wombats 1st	Wombats 2nd	Blue Crabs 1st	Blue Crabs 2nd	Cheater Meters 1st	Cheater Meters 2nd	Sharebots 1st	Sharebots 2nd
Booklet	(1s&5s)	(2s&6s)	(3s&7s)	(4s&8s)	(2s&4s)	(1s&3s)	(6s&8s)	(5s&7s)
Sample Size (n)	337	356	341	351	348	333	360	345
0	**0** 0.00%	**1** 0.28%	**2** 0.59%	**0** 0.00%	**0** 0.00%	**3** 0.90%	**4** 1.11%	**3** 0.87%
1	**1** 0.30%	**6** 1.69%	**3** 0.88%	**7** 1.99%	**2** 0.57%	**4** 1.20%	**22** 6.11%	**22** 6.38%
2	**13** 3.86%	**11** 3.09%	**5** 1.47%	**14** 3.99%	**8** 2.30%	**5** 1.50%	**39** 10.83%	**38** 11.01%
3	**17** 5.04%	**20** 5.62%	**10** 2.93%	**15** 4.27%	**7** 2.01%	**14** 4.20%	**38** 10.38%	**49** 14.20%
4	**32** 9.50%	**40** 11.24%	**21** 6.16%	**33** 9.40%	**17** 4.89%	**16** 4.80%	**64** 17.78%	**68** 19.71%
5	**60** 17.80%	**81** 22.75%	**45** 13.20%	**57** 16.24%	**40** 11.49%	**27** 8.11%	**66** 18.33%	**69** 20.00%
6	**136** 40.36%	**114** 32.02%	**92** 26.98%	**107** 30.48%	**67** 19.25%	**60** 18.02%	**67** 18.61%	**60** 17.39%
7	**78** 23.15%	**83** 23.31%	**163** 47.80%	**118** 33.62%	**71** 20.40%	**77** 23.12%	**41** 11.39%	**20** 5.80%
8	*******	*******	*******	*******	**92** 26.44%	**83** 24.92%	**16** 4.44%	**12** 3.48%
9	*******	*******	*******	*******	**44** 12.64%	**44** 13.21%	**3** 0.83%	**4** 1.16%
Mean	5.54	5.39	6.00	5.58	6.72	6.65	4.56	4.33
Median	6	6	6	6	7	7	5	4

******* : NA

2. Test reliability

Test reliability is one of the important test characteristics. Table 2 shows the reliability for each of the four passages both using raw scores and standardized scores. Generally the reliabilities are higher for standardized scores. The Wombats passage has the lowest reliability; but the reliabilities for all four passages are between 0.52 and 0.61for raw scores, and between 0.55 and 0.64 for standardized scores. It is important to note that reliabilities shown in Table 2 were all based on manifest variables and were influenced by the test length. Generally longer tests have

higher reliability. As all the four passages were short (only 7-8 items for each passage), all of the four passages have a reasonable reliability, adequate for inclusion in a combined passage baseline test.

Table 2: Reliabilities (Cronbach's alpha) of each of the four passages when scaled separately using raw scores and standardized scores

Passage	Grade	Reliability(raw score)	Reliability (standardized score)
Wombats	4	0.52	0.55
Blue Crabs	4	0.61	0.62
Cheater Meters	8	0.59	0.64
Sharebots	8	0.57	0.57

Item-total correlations were also calculated using raw scores and standardized scores. For the passage of Blue Crabs, if the final item were to be deleted from the test, the reliabilities based on standardized scores would be slightly improved from 0.619 and 0.622.

3. Construct validation

Exploratory factor analysis was used to confirm the dimensional structure of the offline reading measures. When items can be categorized, it is recommended that factor analyses should be conducted on the matrix of polychoric inter-item correlations rather than on the matrix of product-moment correlations. Thus, Software package *Mplus* (Muthen and Muthen, 2010) was used for the factor analyses. However, exploratory factor analysis based on polychoric correlations has only been implemented in the *Mplus* for complete datasets. So in this case, an exploratory factor analysis had to be conducted for each passage separately.

Table 3 gives the *Mplus* results for an exploratory factor analysis with a *promax* rotation for all four passages. The solution with one factor was confirmed for all four passages.

Table 3: Exploratory factor analyses for the offline reading assessment items

Grade 4				Grade 8			
Wombats	Promax factor loadings	Blue Crabs	Promax factor loadings	Cheater Meters	Promax factor loadings	Sharebots	Promax factor loadings
Item 1		Item 1	0.478	Item 1	0.401	Item 1	0.583
Item 2		Item 2	0.426	Item 2	0.727	Item 2	0.595
Item 3		Item 3	0.659	Item 3	0.655	Item 3	0.508
Item 4		Item 4	0.746	Item 4	0.447	Item 4	0.631
Item 5		Item 5	0.766	Item 5	0.819	Item 5	0.308
Item 6		Item 6	0.711	Item 6	0.797	Item 6	0.453
Item 7		Item 7	0.436	Item 7	0.328	Item 7	0.520
				Item 8	0.419		

4. Combined 4th and 8th Passages and Potential Order Effects (Booklet effects)

The first question might be whether passages at both levels are necessary to distinguish among students throughout the range of their abilities. In theory, the 8th grade passage ws intended to discriminate among higher performing students, presumably at no expense due to inclusion of the 4th grade passage. Analogously, including the 4th grade passage was designed to distinguish among weaker students whose scores on the 8th grade passage could be too low to be meaningful performance measures. The data largely support both these expectations, indicating that both the lower and the higher level Both Pearson correlation coefficients and examination of the score distributions for both high-performing and weak students on each passage separately are used to consider the wisdom of combining items from two grade levels.

As expected, scores for the two passages were correlated, as shown in Table 4. Because so many students' scores clustered at 6 and 7 on the 4th-grade passage, making little or no distinction among students, the correlations were not strong, ranging from 0.444 to 0.535. This indicates the need to combine passages from the two levels spanning the 7th grade.

Table 4: Pearson correlations for 4th- and 8th-grade passage scores (by ordering of 4th / 8th grade passages)

Correlations		Grade 8	
		Cheater Meters	Sharebots
Grade 4 Passage First / Second	Wombats	0.444 / 0.492	0.480 / 0.477
	Blue Crabs	0.459 / 0.535	0.454 / 0.475

Next, consider the conjecture that the the 4th grade NAEP item will affect high-performing students' total scores very little because it will not discriminate among these students. A total of 338 students completed either Booklet 1 or Booklet 2, with subscores for Wombats and for Cheater Meters. Of the 143 high-performing students (a score of 8 or 9 on Cheater Meters), slightly over three-quarters scored either 6 or 7 on Wombats (and only 6% of these students scored 4 or lower). However, for weak students, the 4th grade item does separate performance. For students scoring 0-4 on Cheater Meters, scores range from 1 to 7 on Wombats, with a median of 4. There are 10 students with scores at the low end (1,2), and 8 students at the top of the scale (6,7) with a predominance (21) of students in the middle (3-5). Thus the 4th grade item gives good separation of students at the bottom of the range of Cheater Meter scores.

The difference in difficulty level is substantial between 4th and 8th grade passages. Therefore concern had been expressed that the order of presentation (easier passage first or more difficult passage first) might affect students' performance, either generally or selectively for students at one end of the performance scale or at the other end.

Comparing mean scores for each possible pairing (e.g., Wombats & Cheater Meters) in Table 5.1, shows that in *all* cases the 4th grade passage subscores were lower when the more difficult 8th grade passage came first. Although the individual statistical (two-tailed) t-test results showed a moderately significant (α=0.10) only for the Blue Crabs & Sharebots pairing, the overall significance of order is documented by the uniformity of the results.

Table 5.1: Students' score summary statistics – 4th grade passage Ssores

	Cheater Meters – 2nd	Cheater Meters – 1st	Sharebots – 2nd	Sharebots – 1st
Wombats Scores	N = 163 Median = 6 Mean = 5.60 Std Dev = 1.2405	N = 175 Median = 6 Mean = 5.54 Std Dev = 1.3464	N = 174 Median = 6 Mean = 5,49 Std Dev = 1.3843	N = 181 Median = 6 Mean = 5.24 Std Dev = 1.5070
Blue Crabs Scores	N = 170 Median =6 Mean = 5.91 Std Dev = 1.4009	N = 173 Median = 6 Mean = 5.66 Std Dev = 1.4684	N = 171 Median = 6 Mean = 6.08 Std Dev = 1.2715	N = 179 Median = 6 Mean = 5.51 Std Dev = 1.5771

For the 8th-grade passage subscores, the results for the different orderings were neither so uniform nor as large. None of the differences in mean scores between orders is statistically significant (two-tailed t-test).

Table 5.2: Students' score summary statistics – 8th grade passage scores

	Wombats – 1st	Wombats – 2nd	Blue Crabs – 1st	Blue Crabs – 2nd
Cheater Meters Scores	N = 163 Median = 7 Mean = 6.69 Std Dev = 1.9546	N = 175 Median = 7 Mean = 6.90 Std Dev = 1.6598	N = 170 Median = 7 Mean = 6.61 Std Dev = 1.8046	N = 173 Median = 7 Mean = 6.53 Std Dev = 1.7271
Sharebots Scores	N = 174 Median = 4 Mean = 4.16 Std Dev = 1.8365	N = 181 Median = 5 Mean = 4.47 Std Dev = 2.0400	N = 171 Median = 5 Mean = 4.50 Std Dev = 1.8736	N = 179 Median = 5 Mean = 4.66 Std Dev = 1.8692

To examine possible selective advantage of one ordering, subsets of high-performing students and of weak students were defined based on their scores on one of the passages. Of particular interest are the students above the 90th %-ile and those below the 10th or even 20th %-ile. (Of course with a very small range of scores, division into exact quartiles is not possible.) Then for each subset of students, the distribution of scores on the second passage was compared when the

4th-grade passage came first and when the 4th –grade passage came second(Tables5.3a&b). For example, the 52 students with perfect scores of 9 on Cheater Meters (top 15.4%) constitute a high-performing subset; while the 39 students scoring 0-4 on Cheater Meters constitute a weak-performing group (bottom 11.5%).

Since there was no restriction on time allocation between the two passages, students may have taken the most time on the difficult passage when it came first. An alternative conjecture is that completion of the questions on the first passage, regardless of level, may have provided "training" for addressing the second passage.

In any case, for 4th-grade passages combined with Sharebots, it appears that high-scoring students on Sharebots scored higher on the 4th-grade passage if they worked on the difficult passage first (Table 5.3a).

To examine possible selective advantage of one ordering on students' performance on the more difficult passage, sets of high-performing and of weak-performing students were redefined this time using students' scores on the easier 4th-grade passage. Then the performance of each subset of students on the more difficult 8th-grade passage was analyzed (Tables 5.4 a&b). The results for weak-performing students are mixed.

Table 5.3a: High score subset summary statistics (top* %-ile for 8th grade passage subscore)
*8th grade passage (Cheater Meters) score = 9; Sharebots score = 7-9

	Cheater Meters – 2nd	Cheater Meters – 1st	Sharebots – 2nd	Sharebots – 1st
Wombats Scores	N = 23 Median = 6 Mean = 5.83 Std Dev = 0.9841 Std Err = 0.2052	N = 29 Median = 6 Mean = 6.0690 Std Dev = 0.9232 Std Err = 0.1059	N = 15 Median = 7 Mean = 6.53 Std Dev = 0.5164 Std Err = 0.5333	N = 30 Median = 6 Mean = 6.00 Std Dev = 0.9469 Std Err = 0.1729
Blue Crabs Scores	N = 21 Median = 7 Mean = 6.57 Std Dev = 0.8701 Std Err = 0.1899	N = 15 Median = 7 Mean = 6.67. Std Dev = 0.6325 Std Err = 0.1633	N = 21 Median = 7 Mean = 6.76 Std Dev = 0.5390 Std Err = 0.1176	N = 30 Median = 7 Mean = 6.23 Std Dev = 1.0400 Std Err = 0.1899

Table 5.3b: Low score subset summary statistics (bottom** %-ile for 8th grade passage subscore)
**8th grade passage (Cheater Meters) score = 0-4; Sharebots score=0-1

	Cheater Meters – 2nd	Cheater Meters – 1st	Sharebots – 2nd	Sharebots – 1st
Wombats Scores	N = 24 Median = 4 Mean = 4.42* Std Dev = 1.5581 Std Err = 0.3180	N = 15 Median = 4 Mean = 3.40* Std Dev = 1.5024 Std Err = 0.3879	N = 16 Median = 3 Mean = 4.00 Std Dev = 1.2500 Std Err = 0.3125	N = 16 Median = 4 Mean = 3.88 Std Dev = 1.6279 Std Err = 0.4070
Blue Crabs Scores	N = 18 Median = 4.5 Mean = 4.11 Std Dev = 1.9670 Std Err = 0.4636	N = 19 Median = 4 Mean = 3.63 Std Dev = 1.5709 Std Err = 0.3604	N = 9 Median = 4 Mean = 4.00 Std Dev = 2.3979 Std Err = 0.7993	N = 10 Median = 3.5 Mean = 3.80 Std Dev = 1.229 Std Err = 0.3888

For students scoring very low on Cheater Meters, the disadvantage of confronting Cheater Meters first and Wombats after (mean difference 1.02 points) is statistically significant (a=0.06, two-tailed t-test))

Table 5.4a: High score subset summary statistics (top***-%-ile for 4th grade passage subscore)
***Wombats score=7; Blue Crabs score=7

	Wombats – 1st	Wombats – 2nd	Blue Crabs – 1st	Blue Crabs – 2nd
Cheater Meters Scores	N = 39 Median = 8 Mean = 7.28 Std Dev = 1.337 Std Err = 0.2140	N = 49 Median = 8 Mean = 7.59 Std Dev = 1.0977 Std Err = 0.1568	N = 78 Median = 8 Mean = 7.24* Std Dev = 1.4611 Std Err = 0.1654	N = 24 Median = 6 Mean = 6.17* Std Dev = 0.8164 Std Err = 0.1667
Sharebots Scores	N = 39 Median = 5 Mean = 5.205 Std Dev = 1.7042 Std Err = 0.2729	N = 32 Median = 6 Mean = 5.60 Std Dev = 1.8625 Std Err = 0.2746	N = 85 Median = 5 Mean = 5.27 Std Dev = 1.6285 Std Err = 0.1766	N = 60 Median = 5.5 Mean = 5.48 Std Dev = 1.4900 Std Err = 0.1924

For students scoring very high (perfect score) on Blue Crabs, the advantage of confronting Blue Crabs first and Cheater Meters after (mean difference 1.07 points) is statistically significant (a=0.01, two-tailed t-test))

Table 5.4b: Low score subset summary statistics (bottom****-%-ile for 4th grade passage subscore) ****Wombats score=0-3; Blue Crabs score=0-3

	Wombats – 1st	Wombats – 2nd	Blue Crabs – 1st	Blue Crabs – 2nd
Cheater Meters Scores	N = 11 Median = 5 Mean = 4.00 Std Dev = 2.3664 Std Err = 0.7135	N = 15 Median = 5 Mean = 4.87 Std Dev = 1.9591 Std Err = 0.5058	N = 10 Median = 3.5 Mean = 3.50 Std Dev = 2.3588 Std Err = 0.7491	N = 16 Median = 4.5 Mean = 4.44 Std Dev = 2.097 Std Err = 0.5242
Sharebots Scores	N = 20 Median = 2 Mean = 2.00 Std Dev = 1.2566 Std Err = 0.2810	N = 23 Median = 2 Mean = 2.65 Std Dev = 1.6406 Std Err = 0.3421	N = 10 Median = 2 Mean = 2.30 Std Dev = 1.4181 Std Err = 0.4485	N = 21 Median = 3 Mean = 2.62 Std Dev = 1.3593 Std Err = 0.2966

Alternate Table 5.4b: Low score subset summary statistics (bottom****-%-ile for 4th grade passage subscore) ****Wombats score=0-4; Blue Crabs score=0-4

	Wombats – 1st	Wombats – 2nd	Blue Crabs – 1st	Blue Crabs – 2nd
Cheater Meters Scores	N = 29 Median = 6 Mean = 5.24 Std Dev = 2.3552 Std Err = 0.4373	N = 32 Median = 5 Mean = 5.31 Std Dev = 2.0230 Std Err = 0.3576	N = 24 Median = 6 Mean = 5,25 Std Dev = 2.3820 Std Err = 0.4862	N = 28 Median = 5 Mean = 4.68 Std Dev = 1.9447 Std Err = 0.3675
Sharebots Scores	N = 34 Median = 3 Mean = 2.85 Std Dev = 1.5789 Std Err = 0.2708	N = 46 Median = 2 Mean = 2.7826 Std Dev = 1.8125 Std Err = 0.2672	N = 17 Median = 3 Mean = 2.59* Std Dev = 1.3720 Std Err = 0.3328	N = 42 Median = 3 Mean = 3.38* Std Dev = 1.8735 Std Err = 0.2891

For students scoring low on Blue Crabs, the advantage of confronting Sharebots first and Blue Crabs after (mean difference 1.19 points) is statistically significant (a=0.02, two-tailed t-test)).

Note however that the percentage of students scoring in the low range (0-4, also true for 0-3) is higher when the 8th-grade passage is administered first. For Cheater Meters, 53/333 (=15.9%) scored 0-4 when the 4th-grade passage was first while 60/348 (=17.2%) scored 0-4 when the 8th-grade passage was first. For Sharebots the results are more extreme with 51/ 345 (=14.8%) scoring 0-4 when the 4th-grade passage was first and 86/360 (=23.9%) scored 0-4 when the 8th-grade passage was first.

5. Statistical and Psychometric Analyses of the Reduced Version of the ORM

For the booklet that combined the passages of Blue Crabs and Sharebots, the model with one factor was confirmed; and the promax factor loadings are shown in Table 6. The raw score distribution has no ceiling or floor effect (see Table 7). The reliability for these combined passages (Cronbach's alpha) is 0.70 for raw scores and 0.72 for standardized scores.

Table 7: Score distribution for combined passages of Blue Crabs and Sharebots (Booklets 7 & 8)

Score Distribution		
Score	Frequency	Percent
0	0	0
1	0	0
2	3	0.85
3	5	1.42
4	6	1.70
5	8	2.27
6	12	3.40
7	18	5.10
8	37	10.48
9	24	6.80
10	42	11.90
11	53	15.01
12	64	18.13
13	41	11.61
14	26	7.37
15	12	3.40
16	2	0.57
17	0	0
18	0	0
19	0	0

Table 6: Exploratory factor analysis factor loadings for combined passages of Blue Crabs and Sharebots (Booklets 7 and 8)

Exploratory Factor Analysis	
Item	Promax factor loadings
SB1S	0.583
SB2S	0.473
SB3S	0.509
SB4S	0.539
SB5S	0.251
SB6S	0.467
SB7S	0.612
BC1S	0.499
BC2S	0.479
BC3S	0.626
BC4S	0.751
BC5S	0.733
BC6S	0.754
BC7S	0.411

Discussion

Based on the analyses results, if the current version of offline reading assessment were reduced to two passages, any of the combinations would be acceptable. Of these the best option is the continue Sharebots (8th grade) and Blue Crabs (4th grade), administering Blue Crabs items first and Sharebots items second (e.g., corresponding to Booklet 7 in Table 1). If reduction of the assessment length is desired, it is recommended to delete the last item from Blue Crabs.

The reasoning for selecting Blue Crabs and Sharebots was due, in part, to the need to expand the score distribution to cover a wide range of student performance, thereby to avoid both floor and ceiling effects. In the case of Wombats, Blue Crabs, and Cheater Meters, a ceiling effect means that a significant percentage of students scored high, i.e., at the top of the scale, so that no further distinctions could be made among a sizeable cluster of high-performing students. Sharebots was the only passage of the four to have a bell-curve distribution of student scores, signifying more accurate discrimination among student abilities. Blue Crabs, intended to discriminate among weaker-performing students, similar in distribution to Wombats but has greater reliability, particularly if the final question were removed. In addition the wording and tone of Wombats' items produced derision among some of the 7th-graders as being too juvenile.

When using these passages, there is a mean differential in scoring based on which passage is given first. Lower-scoring students appear to have fared better (i.e. scored higher) on Sharebots when Blue Crabs was administered second; however this is primarily an artifact occurring because the percentage of students scoring low on the 4th-grade passage is much higher (23.9% compared to 14.8%) when the 8th-grade passage is administered first.

For high-scoring students, the picture is somewhat different. For Cheater Meters, the only important distinction is that in combination with Blue Crabs, the scores on the 8th-grade passage were higher when it was preceded by the easier 4th-grade passage. However, when Sharebots was the 8th-grade passage, the number of students scoring 7-9 for this passage was much higher when the 8th-grade passage came first (60/360 = 16.7% compared to 36/345 = 10.4%), regardless of which 4th-grade passage was included. This suggests that the time devoted to the 8th-grade passage may have been longer when it appeared first and this modified performance on the 4th-grade passage as well. This could also indicate the possible influences of attitude and motivation as well as available time. Those who performed poorly on the 8th-grade passage did better when given an easier task first, regardless of the combination of passages, although the individual result was statistically significant only for the Cheater Meters-Wombats combination. It is unclear whether the advantage to confronting the 4th-grade passage first was because the 8th grade passage was discouraging or because the weaker students' attention to the assessment was declining with time. A student capable of scoring highly may have done better when challenged with the 8th-grade passage first, but may have tended to treat the Blue Crabs passage (an attitude generalized to the test as a whole), as not requiring the same concentration and attention or may simply have expended most of their time on the 8th-grade passage.