Supplementary Material

Table S1

Core Reading Program Components by Grade Level

Grade	Oral Language	Phonological Awareness	High Frequency Words	Phonics	Structural Analysis	Spelling	Fluency	Listening Comprehension	Background Knowledge	Vocabulary	Reading Comprehension	Close Reading	Reading-Writing Connections	Writing Process	Grammar
K	X	X	X	X				X	X	X			X	X	X
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3				X		X	X	X	X	X	X	X	X	X	X
4				X		X	X	X	X	X	X	X		X	X

Note. K = Kindergarten

Fidelity Procedures

The district intended for *Wonders* and *WonderWorks* to be implemented as prescribed by the developers, so the checklists included the specific materials and procedures for each lesson component identified. The pullout intervention did not have prescribed materials, so the checklists addressed the lesson components (e.g., interactive writing, guided reading). For both the core instruction and intervention, the focus of the fidelity evaluation was on whether or not the content and procedures were implemented (i.e., yes or no) rather than estimations of the quality with which teachers delivered those lessons. The curricular programs did not define quality of implementation—only content and procedures.

The last author trained a team of seven coders on the coding process. All coders were female, certified educators with prior classroom experience ($M = 11$ years, range of 5-26 years), and no connections to or knowledge of the teachers or interventionists being monitored. Six coders held graduate degrees. The research team assigned each of five coders to a single grade level, based on their teaching experience implemented. Researchers assigned a sixth coder to the pullout intervention because she was trained on the framework. The seventh coder and the last author conducted reliability checks across grade levels to ensure there was no difference in the stringency of individual coders.

The initial 4-hr training addressed the coding procedures and operationalized definitions in the codebook. Coders received copies of the curricular materials relevant to their assigned files, and they participated in practice sessions with example audio files until they achieved 100% reliability. Once they started working, coders met weekly to resolve issues identified in the recordings, iteratively add detail to the codebook definitions, and prevent observer drift by checking how each was marking particularly instances of instructional fidelity. For example, the reliability coders might play a segment of audio and ask the coders to independently complete a

fidelity checklist. The reliability coders then checked the scores. In addition, approximately 23%
of all audio recordings were coded by two independent coders. Across the weekly calibration
meetings and the ongoing reliability checks, interrater agreement was consistently 100%.

For each teacher, the coders scored one randomly selected audio recording from the first
half of the summer program and one from the last half. While listening to the recordings, the
coders followed along with the teacher's manual or other lesson materials. Scoring was
dichotomous (i.e., each criterion was met or not), and we calculated overall teacher fidelity as the
percentage of criteria met. It is important to note that scoring was stringent, so fidelity was more
difficult to achieve with the highly specified core curriculum (i.e., *Wonders* and *WonderWorks*)
than the more flexible instructional framework used in the pullout intervention. For example, a
Wonders phonological awareness activity might have four steps, so skipping just one step would
result in scoring the lesson as lacking fidelity. On the other hand, if an interventionist did a
writing mini-lesson, only a glaring inaccuracy (e.g., inappropriately using the word *where* for
were) would result in scoring the lesson as lacking fidelity.

Table S2

RAPID Reading Subtests Administered in Each Grade

Subtest	Task Description	Kindergarten	Grade 1	Grade 2	Grades 3-5
Phonological Awareness	Blend heard sounds to say each word	X			
Word Reading	Read aloud words displayed on the screen	X	X	X	
Spelling	Spell by typing each word heard			X	
Word Recognition	Select from 3 options each word heard				X
Vocabulary Pairs	Select 2 of 3 pictures (and corresponding heard words) with similar meanings	X	X	X	
Following Directions	Select or move pictures on screen, following heard directions		X	X	
Vocabulary Knowledge	Select from 3 options the word that completes each sentence				X
Syntactic Knowledge	Select from 3 options the sentence or phrase that best completes a short text				X
Reading Comprehension	Read short passages and answer questions				X

Additional Information on Data Analytic Procedures

Equation for the Pre-post Change Scores Model

The unconditional growth model includes the predictor that represents time in the level-1 model; the level-2 and level-3 specifications allow for baseline and trajectory lines to vary (Hox, 2010; Raudenbush & Bryk, 2002; Singer & Willett, 2003). Thus, this model can be expressed as:

Level-1

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(TIME_{ijk} - 1) + \varepsilon_{ijk},$$

Level-2

$$\pi_{0jk} = \gamma_{00k} + \xi_{0jk},$$

$$\pi_{1jk} = \gamma_{10k} + \xi_{1jk},$$

Level-3

$$\gamma_{00k} = \lambda_{000} + u_{00k},$$

$$\gamma_{10k} = \lambda_{100} + u_{10k},$$

where Y_{ijk} is the outcome score for the ith time point for the jth individual in teacher k, λ_{000} is the baseline, and λ_{100} is the change trajectory. This specification of the model accounts for five residual variances. Specifically, it allows individuals to differ in baseline scores and change score trajectories, and it allows the average baseline and average change trajectory to differ by teacher. Effect sizes for this model were computed as mean change differences (Becker, 1988), where the pooled standard deviations were corrected to take into account the cluster effect via the ICCs for each model.

Equation for the Partially Clustered Model

Level-1

$$Y_{ij} = \beta_{0j} + \beta_{1j}(Treat_{ij}) + \varepsilon_{ij},$$

Level-2

$$\beta_{0j} = \gamma_{00} + u_{0j},$$

$$\beta_{1j} = \gamma_{10},$$

where Y_{ij} is the score of the outcome Y for the ith observation in the jth cluster. In this

specification of the model, the treatment group has several clusters while the control group can

be represented as a single group, β_{0j} is the intercept, β_{1j} is the mean difference between the

treatment and the control group, and $Treat_{ij}$ is a dummy variable that is coded as one if the unit

belongs to the treatment group and zero otherwise. The control group contains a single source of

error at level 1, but the treatment group also accounts for a level-2 error (i.e., between-groups

variability). Recall that the estimation of this model includes weights-based propensity scores.

RAPID Descriptive Statistics by Grade Level

Tables S3 (Kindergarten)

Correlations for Kindergarten Scores

		Pretest				Posttest			
		RSP	WRead	VP	PA	RSP	WRead	VP	PA
Pretest	RSP	1.00	0.80	0.14	0.87	0.87	0.69	0.08	0.74
	WRead		1.00	0.18	0.58	0.80	0.83	0.12	0.67
	VP			1.00	0.12	-0.06	-0.05	0.56	0.04
	PA				1.00	0.78	0.56	0.11	0.74
Posttest	RSP					1.00	0.74	0.06	0.87
	WRead						1.00	0.04	0.47
	VP							1.00	0.17
	PA								1.00

Means and Standard Deviations for Kindergarten Scores, Overall and by Group

Outcome	Group	Test Period	Mean	SD	N	Min	Max
RSP	Treatment	Pre	47.5	27.3	78	1	99
		Post	49.9	26.1	62	1	99
	Control	Pre	39.8	26.7	117	1	99
		Post	-	-	-	-	-
	Overall	Pre	43.0	27.2	195	1	99
		Post	49.9	26.1	62	1	99
WRead	Treatment	Pre	282.0	114.7	78	0	425
		Post	282.7	106.8	62	0	559
	Control	Pre	260.8	135.0	117	0	1000
		Post	-	-	-	-	-
	Overall	Pre	269.7	127.0	195	0	1000
		Post	282.7	106.8	62	0	559
VP	Treatment	Pre	386.1	69.1	78	250	579
		Post	405.8	72.0	62	251	638
	Control	Pre	378.9	69.1	117	218	638
		Post	-	-	-	-	-
	Overall	Pre	381.9	69.0	195	218	638
		Post	405.8	72.0	62	251	638
PA	Treatment	Pre	416.5	115.7	78	88	900
		Post	432.1	122.2	62	19	900
	Control	Pre	381.8	122.8	117	0	900
		Post	-	-	121	-	-
	Overall	Pre	396.3	120.8	195	0	900
		Post	432.1	122.2	62	19	900

Notes. RSP = Reading Success Probability; WRead = Word Reading; VP = Vocabulary Pairs; PA = Phonemic Awareness.

Tables S4 (Grade 1)

Correlations for First-Grade Scores

		Pretest				Posttest			
		RSP	WRead	VP	FD	RSP	WRead	VP	FD
Pretest	RSP	1.00	0.69	0.63	0.30	0.54	0.42	0.39	0.27
	WRead		1.00	0.29	0.26	0.42	0.59	0.28	0.19
	VP			1.00	0.27	0.38	0.25	0.41	0.26
	FD				1.00	0.18	0.14	0.24	0.60
Posttest	RSP					1.00	0.72	0.62	0.28
	WRead						1.00	0.22	0.25
	VP							1.00	0.21
	FD								1.00

Means and Standard Deviations for First-Grade Scores, Overall and by Group

Outcome	Group	Test Period	Mean	SD	N	Min	Max
RSP	Treatment	Pre	31.4	28.7	98	1	95
		Post	40.7	30.2	78	1	96
	Control	Pre	27.1	25.2	116	1	96
		Post	24.9	26.8	106	1	99
	Overall	Pre	28.7	26.6	214	1	96
		Post	31.2	29.2	184	1	99
WRead	Treatment	Pre	446.4	92.1	98	115	576
		Post	475.1	75.7	78	204	584
	Control	Pre	426.0	115.7	116	0	576
		Post	404.1	150.1	106	0	1000
	Overall	Pre	433.7	107.6	214	0	576
		Post	432.6	130.2	184	0	1000
VP	Treatment	Pre	453.1	98.4	98	205	704
		Post	470.7	94.4	78	199	798
	Control	Pre	461.2	81.0	116	220	670
		Post	460.8	92.9	106	199	683
	Overall	Pre	458.1	87.8	214	205	704
		Post	464.8	93.4	184	199	798
FD	Treatment	Pre	459.5	98.8	98	131	721
		Post	488.6	99.7	78	63	718
	Control	Pre	437.3	116.8	116	40	722
		Post	470.5	115.9	106	0	1000
	Overall	Pre	445.7	110.7	214	40	722
		Post	477.8	109.8	184	0	1000

Notes. RSP = Reading Success Probability; WRead = Word Reading; VP = Vocabulary Pairs; FD = Following Directions.

Tables S5 (Grade 2)

Correlations for Second-Grade Scores

		Pretest					Posttest				
		RSP	WRead	VP	FD	SP	RSP	WRead	VP	FD	SP
	RSP	1.00	0.50	0.48	0.61	0.80	0.75	0.41	0.37	0.38	0.66
	WRead		1.00	0.27	0.27	0.55	0.48	0.45	0.29	0.24	0.47
Pretest	VP			1.00	0.39	0.20	0.42	0.21	0.58	0.35	0.25
	FD				1.00	0.29	0.45	0.22	0.35	0.51	0.31
	SP					1.00	0.64	0.56	0.17	0.26	0.76
	RSP						1.00	0.46	0.48	0.62	0.82
	WRead							1.00	0.20	0.22	0.55
Posttest	VP								1.00	0.30	0.22
	FD									1.00	0.34
	SP										1.00

Means and Standard Deviations for Second-Grade Scores, Overall and by Group

Outcome	Group	Test Period	Mean	SD	N	Min	Max
RSP	Treatment	Pre	39.1	27.9	116	1	94
		Post	44.7	28.6	91	1	98
	Control	Pre	34.4	28.0	175	1	98
		Post	37.0	28.7	163	1	98
	Overall	Pre	36.0	28.0	291	1	98
		Post	39.7	28.8	254	1	98
WRead	Treatment	Pre	524.3	77.2	116	115	722
		Post	558.6	78.5	91	382	1000
	Control	Pre	519.4	89.0	175	0	706
		Post	541.8	106.1	163	115	1000
	Overall	Pre	521.0	85.2	291	0	722
		Post	547.6	97.6	254	115	1000
VP	Treatment	Pre	538.9	69.4	116	356	747
		Post	562.7	70.7	91	436	740
	Control	Pre	529.9	91.1	175	266	767
		Post	560.8	98.8	163	342	1000
	Overall	Pre	532.9	84.6	291	266	767
		Post	561.4	89.9	254	342	1000
FD	Treatment	Pre	546.5	136.7	116	245	1000
		Post	593.3	115.8	91	170	806
	Control	Pre	525.5	140.7	175	0	1000
		Post	562.2	129.6	163	56	1000
	Overall	Pre	532.4	139.5	291	0	1000
		Post	573.0	125.7	254	56	1000

Outcome	Group	Test Period	Mean	SD	N	Min	Max
SP	Treatment	Pre	564.7	105.3	93	276	789
		Post	563.7	102.1	95	343	778
	Control	Pre	537.1	126.5	190	100	789
		Post	522.3	134.1	179	100	789
	Overall	Pre	546.2	120.5	283	100	789
		Post	536.6	125.3	274	100	789

Notes. RSP = Reading Success Probability; WRead = Word Reading; VP = Vocabulary Pairs; FD = Following Directions; SP = Spelling.

Tables S6 (Grade 3)

Correlations for Third-Grade Scores

		Pretest					Posttest				
		RSP	WRec	VK	SK	RC	RSP	WRec	VK	SK	RC
Pretest	RSP	1.00	0.65	0.42	0.41	0.85	0.60	0.39	0.41	0.36	0.56
	WRec		1.00	0.20	0.22	0.44	0.43	0.39	0.24	0.18	0.38
	VK			1.00	0.19	0.32	0.26	0.14	0.38	0.08	0.24
	SK				1.00	0.39	0.36	0.24	0.20	0.32	0.33
	RC					1.00	0.51	0.30	0.33	0.35	0.52
Posttest	RSP						1.00	0.63	0.42	0.40	0.85
	WRec							1.00	0.26	0.24	0.40
	VK								1.00	0.30	0.34
	SK									1.00	0.36
	RC										1.00

Means and Standard Deviations for Third-Grade Scores, Overall and by Group

Outcome	Group	Test Period	Mean	SD	N	Min	Max
RSP	Treatment	Pre	22.4	26.1	118	1	99
		Post	22.6	26.1	92	1	98
	Control	Pre	26.8	29.5	210	1	99
		Post	27.3	29.4	194	1	99
	Overall	Pre	25.5	28.6	328	1	99
		Post	25.9	28.5	286	1	99
WRec	Treatment	Pre	296.5	145.3	118	0	872
		Post	280.6	114.1	92	0	601
	Control	Pre	308.1	131.8	210	0	1000
		Post	326.6	129.1	194	0	906
	Overall	Pre	304.8	135.7	328	0	1000
		Post	312.9	126.4	286	0	906

Outcome	Group	Test Period	Mean	SD	N	Min	Max
VK	Treatment	Pre	363.4	81.3	118	0	590
		Post	367.0	78.3	92	0	644
	Control	Pre	370.8	68.3	210	0	752
		Post	378.1	67.3	194	0	651
	Overall	Pre	368.7	72.2	328	0	752
		Post	374.7	70.9	286	0	651
SK	Treatment	Pre	316.9	100.0	118	0	501
		Post	334.1	93.9	92	0	544
	Control	Pre	337.6	89.4	210	0	606
		Post	348.3	93.3	194	0	669
	Overall	Pre	331.7	92.9	328	0	606
		Post	344.0	93.5	286	0	669
RC	Treatment	Pre	327.1	45.7	118	215	475
		Post	326.7	60.5	92	0	493
	Control	Pre	330.5	58.2	210	84	533
		Post	332.0	51.2	194	224	554
	Overall	Pre	329.5	54.9	328	84	533
		Post	330.4	54.1	286	0	554

Notes. RSP = Reading Success Probability; WRec = Word Recognition; VK = Vocabulary Knowledge; SK = Syntactic Knowledge; RC = Reading Comprehension. There were a small number of students in each group who were missing pre- and/or posttest scores.

Tables S7 (Grade 4)

Correlations for Fourth-Grade Scores

		Pretest					Posttest				
		RSP	WRec	VK	SK	RC	RSP	WRec	VK	SK	RC
Pretest	RSP	1.00	0.29	0.51	0.42	0.83	0.50	0.21	0.36	0.34	0.43
	WRec		1.00	0.17	0.27	0.36	0.29	0.23	0.22	0.17	0.28
	VK			1.00	0.33	0.37	0.38	0.20	0.48	0.35	0.37
	SK				1.00	0.44	0.36	0.18	0.33	0.41	0.38
	RC					1.00	0.45	0.20	0.33	0.34	0.45
Posttest	RSP						1.00	0.34	0.60	0.45	0.85
	WRec							1.00	0.37	0.25	0.38
	VK								1.00	0.41	0.47
	SK									1.00	0.50
	RC										1.00

Means and Standard Deviations for Fourth-Grade Scores, Overall and by Group

Outcome	Group	Test Period	Mean	SD	N	Min	Max
RSP	Treatment	Pre	13.2	21.9	72	1	99
		Post	24.8	32.9	53	1	99
	Control	Pre	13.7	23.3	205	1	99
		Post	15.7	25.6	202	1	99
	Overall	Pre	13.6	23.0	277	1	99
		Post	17.6	27.4	255	1	99
WRec	Treatment	Pre	384.2	100.0	72	124	600
		Post	380.3	100.1	53	135	895
	Control	Pre	380.7	107.1	205	0	880
		Post	374.1	113.1	202	0	1000
	Overall	Pre	381.4	105.6	277	0	880
		Post	375.4	110.4	255	0	1000
VK	Treatment	Pre	388.7	86.6	72	150	579
		Post	400.8	80.7	53	177	590
	Control	Pre	379.7	94.8	205	0	696
		Post	384.0	108.3	202	0	723
	Overall	Pre	381.5	93.1	277	0	696
		Post	387.5	103.3	255	0	723
SK	Treatment	Pre	410.1	94.6	72	112	581
		Post	437.8	104.0	53	195	700
	Control	Pre	392.5	117.2	205	0	713
		Post	405.7	113.5	202	0	713
	Overall	Pre	396.1	113.0	277	0	713
		Post	412.4	112.2	255	0	713
RC	Treatment	Pre	362.7	63.6	72	252	559
		Post	382.5	79.0	53	260	580
	Control	Pre	354.8	61.4	205	242	569
		Post	357.0	70.5	202	214	693
	Overall	Pre	356.4	61.8	277	242	569
		Post	362.4	73.0	255	214	693

Notes. RSP = Reading Success Probability; WRec = Word Recognition; VK = Vocabulary Knowledge; SK = Syntactic Knowledge; RC = Reading Comprehension.

References

Becker, B.J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41*, 257-278. doi: 10.1111/j.2044-8317.1988.tb00901.x

Hox, J.J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Singer, J.D., & Willett, J.B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.