

# Appendix

## Appendix A1 Study characteristics: Uchikoshi, 2005 (randomized controlled trial)

Characteristic	Description
<b>Study citation</b>	Uchikoshi, Y. (2005). Narrative development in bilingual kindergarteners: Can <i>Arthur</i> help? <i>Developmental Psychology</i> , 41(3), 464–478.
<b>Participants</b>	The study involved 108 kindergarten students (47 girls and 61 boys). Fifty-one children were assigned to watch <i>Arthur</i> ; 57 were assigned to watch <i>Between the Lions</i> . <sup>1</sup> Picture Vocabulary Test scores indicated that, at the beginning of the intervention, participants' average English vocabulary was at the three-year two-month age level of a monolingual English child. The Spanish version of this measure indicated that their native language vocabulary was at the five-year level; the average age of the children at the beginning of the study was 5 years, 7 months (boys) and 5 years, 6 months (girls). At least 80% of the students in the study qualified for free lunch. The time their families lived in the United States ranged from three months to seven years. According to parent survey responses, only 22% of the children in the sample were born outside of the country. These surveys also indicated that, on average, there were 21 books (in both English and Spanish) in the home, although there was wide variation on this number, ranging from zero to 300.
<b>Setting</b>	The study was conducted in six schools in a large urban district on the East Coast. Spanish-English classrooms (classrooms providing instruction in both languages) were selected, and all teachers were fluent in both languages. All children came from primarily Spanish-speaking homes and neighborhoods with heavy concentrations of Spanish-speaking people.
<b>Intervention</b>	The intervention group watched a 30-minute episode of <i>Arthur</i> at school, three times a week between October and May of one school year, for a total of 54 episodes. Although follow-up activities are available at the PBS website, teachers were directed only to show the videos.
<b>Comparison</b>	The comparison group watched the same number of episodes of <i>Between the Lions</i> over the same time period. <i>Between the Lions</i> is an educational television program with a focus on phonics and reading skills. <i>Arthur</i> focuses on narrative structure. As with the intervention group, none of the follow-up activities associated with the show were used. Each program in this show entails a story that a family of lions read together, focusing on phonological skills and the alphabet.
<b>Primary outcomes and measurement</b>	The outcome measure in the study was an instrument used to assess children's ability to tell a coherent story narrative, total number of words uttered by students, and the average length of the clauses used when describing a story.
<b>Teacher training</b>	Little information about teacher training was provided, other than they were bilingual.

1. Students were assigned within the six classrooms and matched as closely as possible on gender and pretest scores. Each classroom was presumably selected from one school.

## Appendix A2 Outcome measures in the English language development domain

Outcome measure	Description
<b>Combined narrative measure</b>	Students were assessed by asking them to tell a “Bear Story” in English; three pictures of a family of teddy bears served as story prompts. The measure was taken from the School-Home Early Language and Literacy assessment developed by Catherine Snow and colleagues, as cited in Uchikoshi (2005). The measure assesses a student’s ability to develop a coherent narrative in English. Five dimensions are assessed: story structure coding, events coding, evaluation coding, temporality and reference, and storybook language. Coding entailed searching for whether any given dimension is present in the story. So stories were reviewed for an introduction, problem, and resolution (story structure); whether events related to the characters and plot (events coding); whether the children’s perspective were captured in the story (evaluation); presence of temporality and the presence of quotes; and use of adverbs and conjoined noun/verb phrases (storybook language) (Uchikoshi, 2005, p. 468). Children’s narratives were transcribed by trained assessors, and stories were read back to children to ensure they were accurately recorded. Although Spanish outcomes are available, these fall outside the parameters of this review.
<b>Total number of words</b>	The total number of words uttered by students offers a measure of story length.
<b>Mean clause length</b>	The complexity of clauses is thought to be associated with narrative skill development, and the length of clauses served as a proxy. Mean clause length was determined by total number of words (the above measure) divided by the number of clauses.

## Appendix A3 Summary of study findings included in the rating for the English language development domain<sup>1</sup>

Outcome measure	Study sample	Sample size (students)	Author's findings from the study					
			Mean outcome (standard deviation <sup>2</sup> )		WWC calculations			
			Arthur group (column 1)	Control group (column 2)	Mean difference <sup>3</sup> (column 1–column 2)	Effect size <sup>4</sup>	Significance <sup>5</sup> (at $\alpha = 0.05$ )	Improvement index <sup>6</sup>
Uchikoshi, 2005 (randomized controlled trial)								
Combined narrative measure	K	102	4.13 (4.35)	2.34 (3.75)	1.79	0.44	ns	+17
Total number of words	K	102	20.74 (39.4)	10.88 (24.7)	9.86	0.30	ns	+12
Mean clause length	K	102	0.54 (1.67)	0.76 (2.08)	–0.22	–0.11	ns	–5
<b>Domain average<sup>7</sup> for English language development</b>						0.29	ns	+11

ns = not statistically significant

1. This appendix reports findings considered for the effectiveness rating and the improvement index.
2. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
3. Change scores on each measure were calculated (the difference between measures taken in October and May/June) and represent mean scores for the intervention and comparison groups. This difference focuses on how much higher the intervention group scored relative to the comparison condition. This differs from the study author's focus, which was based on how much faster intervention students learned relative to the comparison students. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
4. For an explanation of the effect size calculation, please see the [Technical Details of WWC-Conducted Computations](#).
5. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups. The level of statistical significance was calculated by the WWC and corrects for multiple comparisons. For an explanation see the [WWC Tutorial on Mismatch](#). See the [Technical Details of WWC-Conducted Computations](#) for the formulas the WWC used to calculate statistical significance. These significance levels differ from those in the original study report, because the author presented a growth curve model, which is meant to examine the rate of change among individuals over time. The effect size estimations presented here focus on comparing the rate of change between groups while considering the multiple outcomes (total number of words, mean clause length, and combined narrative measure), thus impacting estimates of whether the groups have statistically significant differences. Note that the study tested outcomes at three time points (October, February, and May/June of the same school year). The WWC analysis used the October and May/June tests as the pre and posttests.
6. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting favorable results.
7. This row provides the study average, which is also the domain average in this case. The WWC-computed domain average effect size is a simple average rounded to two decimal places. The domain improvement index is calculated from the average effect size.

## Appendix A4 Rating for the English language development domain

The WWC rates interventions as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.<sup>1</sup>

For the outcome domain of English language development, the WWC rated *Arthur* as having potentially positive effects. It did not meet the criteria for positive effects, because it only had one study. The remaining ratings (mixed effects, no discernible effects, potentially negative effects, and negative effects) were not considered because *Arthur* was assigned the highest applicable rating.

### Rating received

**Potentially positive effects:** Evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect, thus qualifying as a *positive* effect.  
**Met.** In the one study on *Arthur* that examined English language development, the average effect size was substantively important.
- Criterion 2: No studies showing a statistically significant or substantively important *negative* effect. Fewer or the same number of studies showing *indeterminate* effects than showing statistically significant or substantively important *positive* effects.

**Met.** The WWC analysis found no statistically significant or substantively important negative effects or indeterminate effects in this domain.

### Other ratings considered

**Positive effects:** Strong evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *positive* effects, at least one of which met WWC evidence standards for a strong design.  
**Not met.** *Arthur* had only one study meeting WWC evidence standards. Although the effect was substantively important, the study lacked a strong design.
- Criterion 2: No studies showing statistically significant or substantively important *negative* effects.

**Met.** The WWC analysis found no statistically significant or substantively important negative effects in this domain.

1. For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain level effect. The WWC also considers the size of the domain level effect for ratings of potentially positive effects. See the [WWC Intervention Rating Scheme](#) for a complete description.