Journal of Experimental Child Psychology 187 (2019) 104661

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

journal homepage: www.elsevier.com/locate/jecp

# Self-derivation through memory integration under low surface similarity conditions: The case of multiple languages

Alena G. Esposito[a,*], Patricia J. Bauer[b]

[a] Department of Psychology, Clark University, Worcester, MA 01610, USA
[b] Department of Psychology, Emory University, Atlanta, GA 30322, USA

ARTICLE INFO

*Article history:*
Received 1 March 2019
Revised 1 July 2019
Available online 9 August 2019

*Keywords:*
Knowledge extension
Self-derivation
Elementary children
Semantic memory
Memory integration
Bilingual education

ABSTRACT

A primary objective of development is to build a knowledge base. To accumulate knowledge over time and experiences, learners must engage in productive processes, going beyond what is explicitly given to generate new knowledge. Although these processes are important to accumulating knowledge, they are also easily disrupted. Individuals often depend on surface-level similarities, such as visual features, to recognize the relation between learning episodes. When the surface-level similarity is low, performance on tasks that depend on productive processes, such as self-derivation through integration of new knowledge, suffers. The major purpose of the current research was to examine whether presentation of related information in different languages poses a challenge to memory integration and self-derivation due to low levels of surface similarity between episodes of learning through different languages. In Study 1, 62 children (Grade 2; mean age = 8 years 1 month) listened to story passages containing novel facts that could be integrated to self-derive new knowledge. Related passages were presented either through the same language or through two different languages (cross-language condition; Spanish and English). There were no significant differences between presentation conditions. In Study 2, 100 children (Grades 3 and 4; mean age = 9.7 years) heard novel facts in single sentences, again presented in either a same-language or cross-language condition.

* Corresponding author.
*E-mail address:* aesposito@clarku.edu (A.G. Esposito).

https://doi.org/10.1016/j.jecp.2019.07.001

Whereas third-grade cross-language performance suffered compared with same-language English controls, fourth-grade performance did not. Results suggest that in addition to language proficiency, rich contextual support and experience in a bilingual environment facilitate cross-language integration.


## Introduction

A primary objective of development is to build a knowledge base. To accumulate knowledge over time and experiences, learners must recognize the relevance of one learning episode to another so as to create an integrated representation (Bauer & San Souci, 2010). Moreover, learners must go beyond what is given in learning sessions and use productive processes to generate new knowledge that was not provided through direct tuition. Indeed, learners of all ages capitalize on productive processes, such as deduction and analogy (e.g., Goswami, 1992, 2011; Perret, 2015), allowing for more efficient learning. As examples, children learn mathematical methods to be applied to infinite number combinations and children learn spelling patterns rather than memorizing the correct spelling of each and every word in their language. For this reason, productive processes are assumed to be a major mechanism of cognitive development (e.g., Bauer, 2012; Brown, 1982; Siegler, 1989). Yet, productive processes are also easily disrupted. Individuals often depend on surface-level similarities to recognize that problems or facts are related (e.g., Gentner, 1977). When surface-level similarity is low, individuals struggle to recognize the relation and performance suffers (e.g., Gick & Holyoak, 1983).

Recognizing the relation between episodes is important to the specific productive process that is the subject of the current research, self-derivation of new factual knowledge through integration of separate yet related learning episodes. In self-derivation through integration, recognizing the relation between separate episodes of learning is essential to success (Bauer, King, Larkina, Varga, & White, 2012). Presumably, if children do not recognize the relation between episodes, they will not integrate them and, thus, will not derive new knowledge. Consistent with this premise, children's performance is depressed when the characters in separate yet related story passages are different and, thus, the surface-level similarity of the episodes is lower relative to when the same character is featured and, thus, the surface-level similarity is higher (e.g., Bauer et al., 2012). The major purpose of the current research was to examine how self-derivation through memory integration is affected by a different type of manipulation of surface-level similarity, namely, that resulting from presentation of related information in different languages.

### *Productive processes and surface similarity*

Productive processes are observed across the lifespan and yet are also easily disrupted. Recognizing that content or problems are related when they are presented with few surface-level features in common is a challenge for children as well as adults (e.g., Gentner, 1977; Gick & Holyoak, 1983). For example, young children struggle to recognize numerical equivalency when the surface similarity between groups of objects is low such as three dogs versus three cats (Mix, 1999). As well, across the school-age years, when children are asked to retell a story, they provide less accurate and more inaccurate detail when the characters they use in the retelling are different from the original characters (Gentner & Toupin, 1986). High levels of surface similarity also can be misleading. In the domain of metaphor interpretation, early in development children depend on shared elements to understand metaphors. If the surface similarity of irrelevant elements is high, children have difficulty in ignoring them and, thus, are less likely to appreciate the underlying relational structure (e.g., Gentner, 1988; Winner, Rosenstiel, & Gardner, 1976).

The challenge of navigating conditions of low surface similarity is especially salient in the case of the productive process of analogical problem solving (e.g., Anolli, Antonietti, Crisafulli, & Cantoia, 2001; Brown & Kane, 1988; Day & Goldstone, 2012; Gentner & Rattermann, 1991). For example, Kotovsky and Gentner (1996) examined the development of perceiving relational analogies across childhood. They found that young children were dependent on surface-level features in order to recognize analogous relational structures (see also Brown, Kane, & Echols, 1986, for similar findings). Older children were better able to recognize purely relational commonalities but still had higher performance when surface similarity supported the relational similarity. Across several experiments, they found that recognizing relational similarity is supported not only by surface similarity but also by domain knowledge, language learning, and more experience in making comparisons. Nevertheless, even during adulthood, performance is depressed when surface similarity is low (e.g., Gick & Holyoak, 1980, 1983). Thus, when learning new material and concepts, low levels of surface similarity pose a challenge across the lifespan, especially for young children.

*Self-derivation through integration*

The focus of the current research was the impact of differential surface feature similarity on the specific productive process of self-derivation of new factual knowledge through integration of separate yet related episodes of new learning. The paradigm used to test self-derivation through integration in children begins with the presentation of true but previously unknown facts (i.e., "stem" facts) embedded within richly contextualized story passages. Each story has distinct characters, plots, and settings, and the presentation is separated not only by time but also by other episodes of new learning and buffer activities. Children then are asked questions that can be answered only by generating a novel fact based on the integration of the pairs of related facts (i.e., an "integration" fact). For example, children are presented with a story with the embedded fact "Golden apple seeds taste like almonds" (Stem Fact 1). After a delay, they are presented with another story containing the fact "Apricots are also called golden apples" (Stem Fact 2). The stem facts can be combined to answer the integration question "What do apricot seeds taste like?" (almonds). Performance in a one-stem control condition (children learn only one of the stem facts [but not both] necessary to self-derive) makes clear that both stem facts are necessary for children to produce the integration facts (Bauer & Larkina, 2017; Bauer & San Souci, 2010). Prior research shows a developmental progression in performance. Children as young as 4 years show some success, but the performance of 6- and 8-year-olds indicates steady improvement (open-ended performance of 13%, 50%, and 75%, respectively; Bauer & Larkina, 2017).

Self-derivation through integration is an especially appropriate target for investigation because it is an ecologically valid model of accumulation of knowledge that informs our understanding of how children build knowledge. Four primary findings support this claim. First, research with adults provides evidence that newly self-derived information is rapidly incorporated into the knowledge base. In Bauer and Jackson (2015), based on one 400-ms presentation, adults' event-related potentials (ERPs) to integration facts were intermediate between those to facts that were well known and those to facts that were novel. Based on a second presentation, responses to integration facts became indistinguishable from those to well-known facts and both differed from responses to novel facts. The rapid transition of newly self-derived facts to the status of "well known" suggests that the new information had become incorporated into the knowledge base. Second, the products of self-derivation through integration are retained over time. Indeed, studies of 4- and 6-year-olds reveal virtually no forgetting of newly self-derived knowledge over a 1-week delay (Varga, Stewart, & Bauer, 2016, and Varga & Bauer, 2013, respectively). Third, there is evidence that children in Grades 1–3 (roughly 6–10 years of age) engage in self-derivation through integration in their classrooms (Esposito & Bauer, 2017). Fourth, consistent with the suggestion that self-derivation is a means for accumulating knowledge, elementary school children's self-derivation through integration performance predicts academic achievement in both math and language arts (Esposito & Bauer, 2017); see Varga, Esposito, & Bauer, 2019, for additional evidence in children and consistent evidence in adults). These findings support the contention that the process of self-derivation through integration is an ecologically valid model of accumulation of knowledge.

Most of the conditions of testing self-derivation through integration have revealed it to be a robust process. Yet, there is some evidence that, like other productive processes, it is affected by manipulations of surface similarity. As an illustration, Bauer et al. (2012) tested the effects of different degrees of surface similarity on 6-year-old children's self-derivation of new knowledge through integration. When the characters depicted in the related story passages were the same, performance was high (67% self-derivation in open-ended testing). However, when the characters were different, performance fell (37%). The authors argued that the consistent character served as a cue that the passages were related to one another, thereby facilitating integration. When the characters differed, the cue was absent, making it more difficult for children to realize the relevance of one passage to another. There also is evidence to suggest that self-derivation through memory integration is relatively impervious to manipulations of surface similarity. This evidence comes not from laboratory work but rather from the classroom. Specifically, second-grade children (mean age = 8.17 years) were found to perform equally well under these same low and high surface similarity conditions (36% and 40%, respectively; (Bauer, Esposito, & Daly, 2019). Thus, it remains an open question as to whether low surface similarity conditions present a challenge to self-derivation through memory integration when tested in the classroom.

In the current research, we tested the implications for self-derivation through integration of a different manipulation of surface similarity, namely, that posed when to-be-integrated information is presented through more than one language. This investigation is of specific relevance given that, as of 2016, 22% of children in the U.S. education system spoke a language other than English at home (Kids Count Data Center, 2018) and that a growing number of children are enrolled in bilingual education models (>800 programs as of this writing vs. ~200 in 2000; Center for Applied Linguistics, n.d.). Critically, both of these situations require cross-language integration. Given the challenges associated with the requirement to integrate episodes featuring different characters, it is logical to assume that different languages—with their inherent differences in surface similarity—would impede self-derivation through integration performance. An alternative logical possibility is represented by the interdependence hypothesis, a tenant of which is that learning supports learning regardless of language (Cummins, 1979, 2000). Importantly, evidence supporting this hypothesis is limited to pragmatic literacy skills such as phonemic awareness (e.g., Verhoeven, 1994, 2007). Whether conceptual representations in academic content areas (e.g., science) can be formed through integration of facts introduced through lessons presented in different languages has yet to be tested.

An additional consideration unique to language manipulations is that, unlike the characters in a pair of stories that affect all participants equally, integrating across two different languages could be especially challenging for learners with low proficiency in one language or the other. Work examining the component cognitive processes on which self-derivation depends has consistently found that verbal comprehension predicts performance. In the laboratory, in two parallel studies examining performance in children age 6, 8, and 10 years, verbal comprehension predicted unique variance in self-derivation through integration performance (Esposito & Bauer, 2018). Although reasoning skills were also correlated to performance across both studies, they did not predict unique variance. The consistency in the pattern of findings is especially striking because the two studies used different methods of testing self-derivation through integration (i.e., story passages and individual fact presentation). Even stronger evidence of developmental continuity in the relation between self-derivation through integration and verbal comprehension comes from Varga et al. (2019), in which both college students and elementary school children were tested. Although different paradigms were used, and the adults were tested in the laboratory and children were tested in their classrooms, in both age groups verbal comprehension predicted unique variance in self-derivation performance. The results underscore the importance of accumulated knowledge in the form of verbal comprehension as an important underlying foundation for acquiring new knowledge through self-derivation processes. If children do not yet have a strong foundation in one of the two languages of presentation, performance may suffer.

## *The current study*

In the current research, we examined the impact of differential levels of surface similarity by comparing self-derivation through integration when to-be-integrated material was presented through a single language versus when it was presented through different languages (English and Spanish). The research was made possible through collaboration with a school system offering both a traditional English education program and a Spanish–English bilingual education program (described in the

Appendix A). Through this collaboration, we were able to recruit enough children with the necessary language experiences to examine whether low surface similarity associated with different languages affects self-derivation through integration. We compared the performance of children in the Spanish–English bilingual education program with that of a comparable group of children within the same school in a traditional English education model. Participating children were in elementary grades (Study 1: Grade 2; Study 2: Grades 3 and 4). In addition to self-derivation through integration, we measured verbal comprehension in both English and Spanish, enabling testing of relations between self-derivation through integration and verbal comprehension. The age range is well suited to the question because the children could be expected to have sufficient experience in both languages to complete the task yet also to have sufficient variability in proficiency to allow examination of relations with verbal comprehension. In addition, this is an important age period to study knowledge accrual because it spans the time when children transition from the foundations of literacy and number sense to using these skills to build content knowledge (Center for Public Education, 2015).

We conducted two simultaneous studies using two different methods. For children in Grade 2, novel information was presented through story passages. This is the paradigm of choice for young children because it contains contextual support that is thought to promote successful performance. We used the paradigm developed by Esposito and Bauer (2017), an adaption of the laboratory story passage paradigm that permitted group administration in the classroom. For children in Grades 3 and 4, we used a single-sentence paradigm (Bauer, Blue, Xu, & Esposito, 2016; Esposito & Bauer, 2018). The single-sentence paradigm has the advantage of increasing the number of trials that can be tested. In both studies, we had an English–English control condition both between learning environments (participants) as well as within participants. Both studies also included measures of English and Spanish verbal comprehension to examine relations between language proficiency and self-derivation through integration both within and across languages. We hypothesized that there would be no difference in English–English performance between learning environments (participants in traditional vs. bilingual education). However, for bilingual education participants, we predicted that cross-language performance would suffer compared with English–English performance. We predicted that English verbal comprehension would predict self-derivation performance in the English–English condition and that both Spanish and English proficiency would predict cross-language performance.

**Table 1**
Population demographics.

| | | School system | Study 1 | Study 2 |
|---|---|---|---|---|
| Total *N* | | 1405 | 62 | 100 |
| Racial/ethnic group (%) | | | | |
| | African American | 37 | 13 | 16 |
| | Anglo-American | 33 | 13 | 11 |
| | Hispanic American | 29 | 28 | 52 |
| | Other | 1 | 4 | 10 |
| | Did not report | | 4 | 11 |
| Parent/guardian education (%) | | NA | | |
| | No high school | | 8 | 27 |
| | Some high school | | 5 | 12 |
| | High school | | 13 | 10 |
| | Beyond high school | | 9 | 17 |
| | Technical or associate degree | | 9 | 12 |
| | Bachelor degree or higher | | 11 | 7 |
| | Did not report | | 7 | 15 |
| Home language (*n*) | | NA | | |
| | English | | 31 | 50 |
| | Spanish | | 31 | 50 |

*Note.* NA, not available.

## Study 1

### Method

#### Participants

The participants were 62 second-grade children (34 girls; $M_{age}$ = 8 years 1 month, range = 7;5–8;7 [years;months]). The children were drawn from a school in the rural southeastern United States (see Table 1 for demographic information). Approximately 88% of children in the school system qualify for federally funded school lunch assistance. Consent forms were sent home through parent communication folders (the typical means of communication between the school system and children's parents/guardians). Only the data from children whose parents/guardians returned signed consent forms were included in analyses (~68% of the population).

The children represented in this sample comprise a matched-sample subset from a larger ongoing study ($N$ = 184 second graders). Participants were drawn from the same pool as Esposito and Bauer (2017). However, all of the data for the current report are original to it. Each child in the traditional education program was matched to a child in the bilingual program based on home language exposure (per parent report), nonverbal intelligence, English vocabulary and analogies, and caregiver education level (per parent report) to form a yoked control group. It is important to note that the "matching" process resulted in statistically equivalent groups on all the variables on which the groups were matched. As a result, the variables are eliminated as sources of group differences, although they still are potential sources of individual difference (see Table 2).

Participating parents/guardians were thanked with a $10 gift card to a local merchant, and children were thanked with a small toy (e.g., a ball). Participating school personnel were thanked with a $20 gift card to a local merchant. The university institutional review board and the participating school system school board reviewed and approved all protocols and procedures for this study and the subsequent study.

#### Stimuli

The stimuli were two novel "stem" facts from each of four domains. Within each domain, the two novel stem facts were related and could be combined to generate a novel integration fact (e.g., the "Apricot seeds taste like almonds" example presented above). The facts all were accurate and determined to be novel for children in the target age range. Laboratory testing revealed that both stem facts were necessary for production of the integration facts. That is, when only one of the two stem facts was provided (one-stem condition), children did not generate the integration facts. The same pattern was found in Bauer and Larkina (2017) and Bauer and San Souci (2010), establishing the validity of the paradigm as a test for integration.

**Table 2**
Matched sample description.

| | Panel A: Study 1 | | Panel B: Study 2 | | | |
|---|---|---|---|---|---|---|
| | Grade 2 | | Grade 3 | | Grade 4 | |
| | Single language ($n$ = 31) | Dual language ($n$ = 31) | Single language ($n$ = 22) | Dual language ($n$ = 22) | Single language ($n$ = 28) | Dual language ($n$ = 28) |
| Nonverbal intelligence | 11.06 (1.93) | 11.06 (2.13) | 11.73 (1.83) | 11.77 (1.54) | 12.70 (2.54) | 13.39 (2.18) |
| English verbal comprehension | 28.36 (4.95) | 28.69 (6.55) | 29.67 (3.69) | 27.35 (7.84) | 31.64 (5.18) | 31.25 (6.78) |
| English analogies | 13.32 (2.68) | 13.42 (3.31) | 14.50 (2.28) | 13.72 (4.35) | 15.46 (3.19) | 16.50 (3.68) |
| Caregiver education | 3.69 (1.76) | 3.85 (1.83) | 3.00 (1.84) | 3.53 (1.95) | 2.70 (1.73) | 2.86 (1.70) |
| Spanish verbal comprehension | 10.79 (12.08) | 19.31 (9.72) | 15.83 (14.71) | 26.20 (12.67) | 16.08 (13.76) | 21.75 (11.77) |
| Spanish analogies | 6.50 (7.69) | 8.12 (6.28) | 8.17 (8.99) | 15.55 (6.41) | 8.56 (7.62) | 14.58 (8.92) |

*Note.* All measures are reported as means (and standard deviations).

As in Esposito and Bauer (2017), the stem facts were featured in text passages presented via PowerPoint as digital stories (see Bauer & San Souci, 2010, for an example text passage). The passages were 81–86 words in length (English and Spanish stories both had a mean of 83 words) distributed over four pages. Each page consisted of a hand-drawn illustration depicting the main actions of the text; the text was not featured on the page. The passages were similar in structure; in each passage, a character (e.g., a squirrel) learned a novel fact in the course of a short "adventure." To avoid ceiling effects, based on the results of Esposito and Bauer (2017), each story passage had a different main character (e.g., a squirrel in one text passage about apricots and a butterfly in the other text passage about apricots). Prerecorded audio tracks of the passages were played through speakers.

There were different versions of the stimuli based on children's educational placement (traditional single language vs. bilingual dual language; see Table 3 (Panel A) for a description and example of how stimuli were presented across conditions). In both education programs, there was a control condition and a manipulated condition (within participants). The control condition was the same for both groups of children and consisted of two sets of stimulus pairs, with both stem facts presented in English such that both passages (and thus both stem facts) were read by the same native English speaker. For children in the bilingual program, the manipulated set of stimuli included two sets of stimulus pairs presented across languages. Thus, for each pair of stimuli, one passage (and thus one stem fact) was presented in English by a native English speaker and the other passage (and stem fact) was presented in Spanish by a native Spanish speaker. For children in the traditional education program, the manipulated condition passage pairs all were presented in English but read by different native English speakers. This mimicked the experience of the children in the bilingual program, who heard different speakers across languages in the cross-language condition and allowed us to examine whether a change in speakers between related passages influences performance.

In summary, traditionally educated (single-language) students received four stimulus pairs, for a total of eight stem facts, through English. The control condition was composed of two pairs that were presented by the same speaker, and the manipulated condition was composed of two pairs that were

**Table 3**
Language manipulations by educational program.

| | Stem fact passage | Educational program: Dual language | Single language |
|---|---|---|---|
| *Panel A: Study 1, Grade 2* | | | |
| Control (English–English) (2 stimulus sets) | Stem 1 | English Speaker 1<br>Golden apple seeds taste like almonds. | English Speaker 1<br>Golden apple seeds taste like almonds. |
| | Stem 2 | English Speaker 1<br>Apricots are also called golden apples | English Speaker 1<br>Apricots are also called golden apples |
| Manipulated (2 stimulus sets) | Stem 1 | English Speaker 2<br>Golden apple seeds taste like almonds | English Speaker 2<br>Golden apple seeds taste like almonds |
| | Stem 2 | Spanish Speaker 1<br>Los albaricoques también son llamados manzanas de oro | English Speaker 3<br>Apricots are also called golden apples |
| *Panel B: Study 2, Grades 3 and 4* | | | |
| Control (English–English) (8 stimulus sets) | Stem 1 | English Speaker 1<br>Titan is Saturn's largest moon | English Speaker 1<br>Titan is Saturn's largest moon |
| | Stem 2 | English Speaker 1<br>Saturn's largest moon is the only moon that has clouds | English Speaker 1<br>Saturn's largest moon is the only moon that has clouds |
| Manipulated (8 stimulus sets) | Stem 1 | English Speaker 2<br>Titan is Saturn's largest moon | English Speaker 2<br>Titan is Saturn's largest moon |
| | Stem 2 | Spanish Speaker 1<br>La luna más grande de Saturno es la única con nubes | English Speaker 3<br>Saturn's largest moon is the only moon that has clouds |

presented by different English speakers. Children in the bilingual education program also received four stimulus pairs. The control condition was the same, with two pairs (and thus four stem facts) presented through English. The other two pairs were presented in a cross-language condition, with one stem fact from each pair presented through English and one through Spanish, for a total of two English stem facts and two Spanish stem facts in this condition. Thus, both groups had four story pairs and eight stem facts across the two conditions, but the bilingual education participants heard six stem facts through English and two through Spanish.

*Measures*

For each child whose parents provided consent, the school provided demographic information on the child's sex, birthdate, and race/ethnicity. Parents/guardians completed a questionnaire (91% compliance), including information on language use and caregiver education, a critical socioeconomic component for development (e.g., Hoff, 2013).

We tested nonverbal intelligence in a group presentation. In this task, individuals choose one of four or five images to complete a pattern. Images were projected via Turning Point software, and children recorded their answers with an individual response device (a.k.a., a "clicker"). Two examples were presented to the class as a whole; corrective feedback was provided. Children then were presented with 15 items (the number of test questions presented was based on individual testing within the same community the prior year). Children were not provided feedback on the test items. The raw number of correct items was recorded as the outcome variable.

The Woodcock–Muñoz Language Survey–Revised Normative Update (WMLS–RNU) is a norm-referenced measure of verbal comprehension, available in both English and Spanish, and is appropriate for ages 2–90+ years. We used both English and Spanish language comprehension measures: Verbal Comprehension Test 1, vocabulary, and Test 2, analogies; both tests were administered in both languages. Raw scores within each language were recorded and summed for a total of two verbal comprehension measures, one in English and one in Spanish. Descriptive statistics are provided in Table 2 (Panel A).

*Procedure*

There were two sessions; Session 1 was group administered in the classroom, and Session 2 was individually administered. Both sessions took place three quarters of the way through the academic year (i.e., in February and March). The first author and three research assistants completed Session 1. Researchers formed teams of two for each classroom presentation. The 45-min classroom sessions were divided into three phases: (a) exposure to the first set of previously unknown stem facts, (b) exposure to the second set of previously unknown stem facts, and (c) test for self-derivation of new factual knowledge through integration of pairs of related stem facts.

In Phase 1 (exposure to the first set of true but previously unknown facts), children heard one text passage from each of four stimulus domains (e.g., "Golden apple seeds taste like almonds"). Illustrations conveying the main actions of the passages were projected onto a screen (~4 by 6 feet). The prerecorded audio tracks were played through speakers. After exposure to four stem fact passages, children engaged in a buffer activity with individual papers at their desks. The buffer activity took approximately 10 min.

Phase 2 commenced after the buffer activity. Children heard the second member of each stem fact story pair, one from each of the four stimulus domains (e.g., "Apricots are also called golden apples" or "Los albaricoques también son llamados manzanas de oro" for bilingual education children). For both Phase 1 and Phase 2, the slides and audio were advanced automatically, ensuring consistent timing across classrooms. Children then engaged in the group-administered nonverbal intelligence task, which took approximately 10 min.

Phase 3 tested for derivation of new factual knowledge and recall of stem facts. All questions were read aloud while researchers monitored. Children began with open-ended paper-and-pencil responses testing self-derivation of integration facts (e.g., "What do apricot seeds taste like?") and recall of the individual stem facts (e.g., "What is a golden apple?"). Children then used their response devices to answer the same integration and stem questions in a forced-choice format in that order. Forced-choice testing included three answer choices. We included both open-ended and forced-choice formats to increase the likelihood that we would have the necessary variance for analyses. Forced-choice performance tends to be higher than open-ended performance (e.g., Bauer & Larkina, 2017). Thus, if performance was poor in the manipulated open-ended condition in particular, we had the possibility of greater variability from the forced-choice measures. Note that during Phase 3 children did not have access to the stem facts. They were required to rely on their memory representations for relevant information.

The stimulus sets and stem order (1 and 2) were counterbalanced such that each stimulus set was presented by one speaker and two speakers equally often and members of stem fact passage pairs were read first and second equally often. In the bilingual classrooms, each stimulus set was used equally often in the English–English and English–Spanish conditions. In the English–Spanish condition, the Spanish story passage was presented second (to block all Spanish language stories rather than alternating between languages within a phase). Within the bilingual program, presentation of material while in the Spanish or English learning environment was counterbalanced such that half of the children tested were in their Spanish classroom and the other half were in their English classroom. The text passages within domains were counterbalanced, and domains were presented in one of four predetermined random orders; each order was used approximately equally often across classrooms. The integration and stem fact questions were presented in one of four random orders; each was used roughly equally often across classrooms and text passage orders.

Session 2 took place 1 week after Session 1 (mean delay = 7.14 days) in a quiet classroom provided by the school. Children provided assent before participating. As part of a larger study, all children completed tests of both Spanish and English verbal comprehension in that order. The order was chosen to ensure that children would be able to end the session having been successful on a task; all children had some knowledge of English due to classroom instruction, but not all children had knowledge of Spanish (native English speakers in traditional classrooms). All children completed the same battery, ensuring that testers were unaware of classroom assignment and ensuring that no assumptions were made regarding language proficiency. All individual testing was conducted by one of eight bilingual research assistants (either native Spanish speakers or Spanish students who completed advanced-level courses, including a study abroad in a Spanish-speaking country). Research assistants were extensively trained and were monitored by the first author during data collection to ensure protocol fidelity.

*Scoring*

Children received scores for integration facts (self-derived) and for stem facts (directly taught). For the integration facts, children received scores in both the control and manipulated conditions in both open-ended and forced-choice formats. The children in traditional education received open-ended and forced-choice scores for stimuli (all in English) presented by the same speaker (control; max = 2) and for stimuli presented by two different speakers (manipulated; max = 2). Children in the bilingual program received a score for English-only material presented by the same speaker (control; max = 2) and a score for cross-language presented material presented by two different speakers and through two different languages (manipulated; max = 2). For the stem facts, children in traditional education received 1 point for each correctly recalled stem fact in open-ended (max = 8) and forced-choice (max = 8) formats. Children in the bilingual program, who received two pairs through English and two pairs in a cross-language condition, were given an English stem score (English max = 6) and a Spanish stem score (Spanish max = 2) in both open-ended and forced-choice formats. To control for differences in the number of stem facts presented in each language, we then converted scores to percentages.

*Data analyses*

We first examined group differences in self-derivation through integration performance by educational program and condition. We analyzed open-ended (recall) performance first in a mixed-factor analysis of variance (ANOVA) with education program (traditional or bilingual) as a between-participants factor and condition (control or manipulated) as a within-participant factor. Before examining group differences in forced-choice performance, we first determined whether performance was above chance with a *t* test. We then conducted the same mixed-factor ANOVA. We then repeated the procedure to examine education program and condition differences in stem fact performance.

Second, we examined individual differences in self-derivation performance. We conducted regression analyses to examine the unique predictive power of English verbal comprehension and Spanish verbal comprehension. There were two models. In Model 1, to test the hypothesis that English verbal comprehension would predict performance in English, we examined total performance for all participants across education models and language conditions. In Model 2, we specifically tested the hypothesis that both English and Spanish verbal comprehension would predict performance in the cross-language condition by examining performance of the cross-language stimuli (only the bilingual education participants contributed data).

## *Results*

The results are reported in three sections. First, we examined children's self-derivation of new factual knowledge through integration of separate yet related episodes of new learning. Second, we examined recall and forced-choice recognition of the stem facts. Third, we examined English and Spanish verbal comprehension as predictors of successful self-derivation through integration. All analyses were conducted using the SPSS Statistics package (Version 24). All statistical tests were two-tailed. Bonferroni corrections were made where appropriate.

### *Self-derivation through integration*

The mean numbers of novel integration facts produced are provided in Table 4 by educational program and language condition. The overwhelming proportion of incorrect responses in open-ended testing stemmed from children marking "I do not know," leaving too few commission errors for analysis. To examine main effects and a possible interaction, we conducted mixed-factor ANOVAs with education program (traditional or bilingual) as a between-participants factor and condition (control or manipulated) as a within-participant factor. Open-ended performance and forced-choice performance were examined in separate analyses. The dependent measures are reported in Table 4.

**Table 4**
Integration and stem fact performance.

| *Panel A: Study 1, Grade 2* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Integration [mean (*SD*)] | | | | Stem facts [proportion (*SD*)] | | | |
| | | Open-ended | | Forced choice | | Open-ended | | Forced choice | |
| | | Control | Manipulated | Control | Manipulated | English | Spanish | English | Spanish |
| Grade 2 | Single language | .60 (.77) | .67 (.61) | 1.61 (0.57) | 1.61 (0.69) | .56 (.21) | NA | .78 (.26) | NA |
| | Dual language | .43 (.63) | .70 (.60) | 1.52 (0.63) | 1.52 (0.68) | .50 (.25) | .57 (.31) | .87 (.19) | .74 (.36) |
| *Panel B: Study 2, Grades 3 and 4* | | | | | | | | | |
| | | Integration [proportion (*SD*)] | | | | | | | |
| Grade 3 | Single language | .07 (.09) | .14 (.16) | .49 (.22) | .48 (.19) | | | | |
| | Dual language | .09 (.09) | .05 (.10) | .54 (.16) | .36 (.17) | | | | |
| Grade 4 | Single language | .17 (.16) | .20 (.23) | .52 (.28) | .53 (.25) | | | | |
| | Dual language | .13 (.15) | .12 (.14) | .54 (.22) | .54 (.25) | | | | |

*Note.* For Grade 2, integration means come from 2 trials; thus, the maximum score is 2. Stem fact percentages are based on either 2, 6, or 8 trials, depending on the language. For Grades 3 and 4, integration percentages are based on either 7 or 8 trials. NA, not applicable.

Self-derivation through integration means are based on number correct and range from 0 to 2 in each condition (2 trials per condition).

Analysis of children's open-ended performance revealed no significant main effects ($Fs < 2.56$, $ps > .12$) and no interaction ($F = 0.92$, $p = .34$). Across education models and conditions, children self-derived novel integration facts on approximately 30% of the trials. Performance was higher in forced-choice testing, with children selecting the correct response alternative on approximately 77% of trials; forced-choice performance was reliably greater than chance (33%), $t(61) = 12.89$, $p < .001$, $d = 1.64$. Similarly, the mixed-factor ANOVA of forced-choice performance revealed no significant main effects ($Fs < 0.32$, $ps > .58$) and no interaction ($F = 0.04$, $p = .84$). Thus, children showed evidence of self-derivation through integration across all conditions in both open-ended and forced-choice formats. The manipulated conditions did not impede performance for children participating in either manipulation (English presentation with a change in speaker or cross-language presentation with changes in both language of presentation and speaker). Children participating in the bilingual education program did as well as their peers across all conditions, evidencing no impediment even when they were required to integrate information across languages.

*Stem facts*

Children's levels of recall and forced-choice recognition of the stem facts are shown in Table 4 by educational model and language condition. Children in the traditional education program had the opportunity to answer eight stem fact questions on content provided through English. In contrast, children in the bilingual program had the opportunity to answer six stem fact questions with content provided through English and two with content provided through Spanish. Thus, we report percentage correct rather than number correct. We conducted separate analyses (a) across education models to compare recall and forced-choice recognition of stem facts presented in English and (b) within the bilingual education model to compare recall and forced-choice recognition of stem facts presented in English versus Spanish. Across education models and languages, forced-choice stem fact performance was reliably above chance (33%), $t(61) = 15.54$, $p < .001$, $d = 1.97$.

We analyzed accuracy rather than raw scores to evaluate recall and forced-choice recognition of stem facts between education models for content presented in English, allowing for comparison even though the possible range of the outcome measures differed. Neither open-ended format, $t(58) = 1.05$, $p = .30$, $d = 0.27$, nor forced-choice format, $t(60) = -1.44$, $p = .15$, $d = 0.37$, differed significantly by educational model. For the children in the bilingual program, we again used accuracy to account for differences in the possible range of the measures to compare recall and forced-choice recognition of stem facts presented in English versus Spanish. The paired-sample $t$ test revealed no differences in performance in open-ended format, $t(29) = -1.28$, $p = .21$, $d = 0.23$, but forced-choice format differed such that English performance was higher than Spanish performance, $t(30) = 3.08$, $p = .004$, $d = 0.55$.

In summary, children recalled roughly one half and three quarters of the stem fact questions in open-ended and forced-choice formats, respectively. In forced-choice testing, frequency of selection of the correct alternative was reliably greater than chance. Stem fact performance on English facts did not differ between education models. Performance on stem facts between languages (tested with children in the bilingual program only) did not differ in open-ended format, but in forced-choice format children had higher performance on English stem facts than on Spanish stem facts. Forced-choice stem fact performance was high across education models and conditions, with performance over 70% even in the Spanish forced-choice condition.

*Verbal comprehension*

To examine possible relations between self-derivation through integration and verbal comprehension, we tested regression models with verbal comprehension as a predictor of self-derivation performance. We examined the influence of both English and Spanish across all conditions because there were native speakers of both languages across all conditions. Thus, it was important to test whether accumulated knowledge as measured by verbal comprehension was related to self-derivation performance regardless of the language through which information was presented. Results are reported in Table 5 (Panel A). We conducted analyses predicting total self-derivation performance across education models and language conditions as well as a model predicting only cross-language integration

**Table 5**
Regression models.

| | | Panel A | | Panel B | |
|---|---|---|---|---|---|
| | | Model 1: Total performance (all participants) | Model 2: Cross-language performance (dual-language participants) | Model 1: Total performance (all participants) | Model 2: Cross-language performance (dual-language participants) |
| Predictor | | $\beta$ | $B$ | $\beta$ | $\beta$ |
| Grade | | NA | NA | .19 | .38* |
| | $R^2$ | | | .04 | .14* |
| Grade | | NA | NA | .12 | .37* |
| English verbal comprehension | | .59* | .63* | .26* | .16 |
| Spanish verbal comprehension | | .27 | .24 | −.007 | .31* |
| | $R^2$ | .23** | .25* | .10* | .23* |

*Note.* All coefficients listed are standardized ($\beta$). NA, not applicable.
* $p < .05$.
** $p < .06$.

performance. Both the model predicting total self-derivation (full sample) and the model predicting cross-language self-derivation (bilingual program participants only) were significant. Across both models, only English verbal comprehension was a significant predictor. Those higher in English proficiency outperformed their lower-proficiency peers in self-derivation. Importantly to the hypotheses, in the cross-language condition, only English verbal comprehension predicted performance even though half of the stem facts were presented in Spanish.

## *Discussion*

The results of this study are counter to our predictions that the low surface similarity of episodes presented through different languages would impede self-derivation through integration performance. Performance in the cross-language condition did not statistically differ from single-language integration, indicating that children were able to navigate integration across two different languages and two different speakers with the same success as integrating within a language with only one speaker. We also expected that cross-language integration specifically would be especially difficult for children with less second-language proficiency. This prediction was partially supported in that verbal comprehension related to self-derivation performance, but the expectation that Spanish verbal comprehension would predict cross-language self-derivation performance was not supported. Overall, children self-derived new knowledge through integration across all control conditions and language/speaker manipulations.

Stem fact performance was also high, with few differences between languages of presentation. There were no differences in open-ended recall of the stem facts between education models. Forced-choice performance differed only within the language manipulation such that stem facts presented through English were recognized at a higher rate than stem facts presented through Spanish, although the effect was small and performance was high in both conditions. This could also be due to a difference in the number of trials such that missing one Spanish stem fact question resulted in a percentage score of only 50%. Overall, performance was high, especially in forced-choice format, indicating that the children had encoded the presented content.

The most significant predictor of self-derivation performance was English verbal comprehension even for the cross-language condition. This was counter to our prediction that the cross-language condition would also depend on Spanish verbal comprehension such that knowledge in both languages would be necessary to integrate across them. The results are consistent with the interdependence hypothesis, which posits that learning supports learning regardless of language (Cummins, 1979, 2000). Indeed, there is evidence that children have access to concepts through either language from early in the language learning process (Poarch, Van Hell, & Kroll, 2015). This could suggest that integrating across languages poses little challenge to even early language learners.

Alternatively, the rich contextual support provided within the stories may have mitigated the impact of the low surface similarity cross-language presentation. Information was presented through context-rich story passages that featured characters and plots and were supported by illustrations. These features may have been sufficient to cue children to the relations between the members of the pairs of story passages even when they were presented through different languages. In addition, there are two aspects of the method that likely contributed to the importance of English for cross-language performance. By design, the first stem fact story passage always was provided through English, meaning that first exposure to the domain was through English. In addition, all questions were asked through English. Future research should examine whether these design decisions affect performance.

Whereas the paradigm used in the current study had advantages, it also permitted only a few test trials and thus a restricted range in the dependent variable (0–2). This limitation could not be overcome with additional trials in this age group out of concern for depression of performance. Yet for older children (Grades 3 and 4), we were able to use a different parallel paradigm that allows additional trials (Esposito & Bauer, 2018). Prior laboratory research in a single-sentence format, without attendant stories, characters, plots, or illustrations, has shown self-derivation through integration in children aged 7–10 years (Bauer et al., 2016; Bauer, Dugan, Varga, & Riggins, 2019) as well as test–retest reliability (Esposito & Bauer, 2018). In Study 2, we used the single-sentence paradigm to examine self-derivation performance by children in Grades 3 and 4, extending the grades previously tested in the classroom to include Grade 4 (Esposito & Bauer, 2017). The single-sentence paradigm quadrupled the number of trials we were able to administer.

## Study 2

### *Method*

#### *Participants*

The participants were 100 children (58 girls) across Grade 3 ($n = 44$; $M_{age} = 9$ years 2 months, range = 8;6–10;1) and Grade 4 ($n = 56$; $M_{age} = 10$ years 1 month, range = 9;3–10;9). They were drawn from the same source and represent the same population as Study 1. None of the children who participated in Study 1 took part in Study 2. The children comprise a matched-sample subset from Grades 3 and 4 of the larger study ($N = 325$ across third and fourth grades). The matched sample was created using the same criteria as Study 1. See Table 1 for sample description. Participating children had been in their education program (traditional or bilingual) since school entry. Parents, teachers, and participating children were again thanked with gift cards and small toys.

#### *Stimuli*

The stimuli were 32 novel "stem" facts that could be integrated to create 16 novel "integration" facts. Preliminary testing revealed that both stem facts were necessary for production of the integration facts and facts were novel to children in the target age range.

As reflected in Table 3 (Panel B), following the procedure for Grade 2, there were different conditions of the stimuli based on education program. Children in both education programs participated in an English–English control condition with 8 fact pairs (16 individual facts) presented through English. In the manipulated condition, the children in the bilingual program were tested on 8 fact pairs in cross-language format in which one fact was presented through English and the other related fact was presented through Spanish (total of 8 stem facts presented in English and 8 stem facts presented in Spanish). Children in the traditional education program had a manipulated condition with 8 fact pairs in which one fact was presented by one English speaker and the related fact was presented by a different English speaker for a total of 16 stem facts presented in English. All facts were presented by a native speaker of the language in which they were presented.

*Measures*

The measures, nonverbal intelligence and verbal comprehension, were the same as in Study 1, including school- and parent/guardian-provided demographic information. In the nonverbal intelligence measure, children in Grade 3 were presented with 16 items and children in Grade 4 were presented with 18 items (based on individual testing within the same community the prior year). See Table 2 (Panel B) for descriptive statistics.

*Procedure*

The procedure resembled that for Study 1 and took place three quarters of the way through the academic year (in February and March). Session 1 was group administered in the classroom, and Session 2 was individually administered. Session 1 was administered by the same four researchers in two-person research teams. The session took roughly 45 min and consisted of the same phases as Study 1: (a) exposure to the first set of stem facts, (b) exposure to the second set of stem facts, and (c) test for self-derivation of new factual knowledge through integration of pairs of related stem facts. Children recorded their forced-choice responses using Turning Point software and individual response devices. Facts were projected onto a screen (~4 by 6 feet). Audio tracks were prerecorded and played through speakers.

In Session 1, Phase 1 and Phase 2 exposed children to the first and second sets of stem facts, respectively. The stem facts were presented in one of four predetermined random orders; each order was used approximately equally often across classrooms. The facts also were counterbalanced such that across classrooms they were used in the control and manipulated conditions equally often. Each fact was presented as the first and second stem fact equally often. Within the bilingual program, the language in which the first stem fact was presented was counterbalanced such that each class received half of the facts in each order. Facts were presented in blocks of four separated by a short nonverbal video clip. Language remained consistent within each block.

In Phase 1, children saw and heard 16 individual unrelated stem facts. To encourage engagement during exposure without the story and illustrations of Study 1, we incorporated categorization engagement questions following half of the facts (8 randomly interspersed). For example, after seeing and hearing "Titan is Saturn's largest moon," children were asked "Was this fact about planets or animals?", to which they responded with the clicker. The engagement questions were asked in the same language as the facts to which they referred. After exposure to 16 stem facts, children completed the same buffer activity as in Study 1.

Phase 2 commenced after the buffer activity. Children saw and heard the second member of each stem fact pair (16 facts). Engagement questions were again interspersed among the stem facts. For both Phase 1 and Phase 2, the slides and audio were advanced automatically, ensuring consistent timing across classrooms. Following Phase 2, children engaged in the group-administered nonverbal intelligence measure for approximately 10 min.

Phase 3 tested for self-derivation of new factual knowledge through integration in open-ended and forced-choice formats. To reduce participant burden, children were not tested for recall of the individual stem facts. Due to a procedural error, one stem fact question was posed instead of its attendant integration question, which was omitted from analyses. Performance on the stem fact question was not included in analyses. The integration questions were first tested in open-ended format via paper handout. Each child had one of four versions of the same questions presented in a different random order; children sitting next to each other received different versions. After open-ended testing, the same integration questions were presented in forced-choice format via Turning Point. Again, questions were presented in one of four predetermined random orders; each order was used approximately equally often across classrooms and stimulus orders. Children self-read the questions while two researchers circulated around the classroom.

Session 2 followed the same procedure as in Study 1. The session took place approximately 1 week after Session 1 (mean delay = 7 days) and took approximately 45 min.

*Scoring*

Children received self-derivation performance scores in both the control and manipulated conditions in both open-ended and forced-choice formats. The children in the traditional education program received open-ended and forced-choice scores for stimuli presented in English by the same speaker (control; max = 8) and stimuli presented in English by two different speakers (manipulated; max = 8). Children in the bilingual program received a score for English-only material presented by the same speaker (control; max = 8) and a score for cross-language (Spanish–English) material presented by two different speakers (manipulated; max = 8). To control for differences in the number of integration questions across conditions (due to the procedural error that resulted in 15 integration questions rather than 16), we then converted scores to percentages.

*Data analyses*

We first examined performance on self-derivation through integration performance. Open-ended performance and forced-choice performance were compared with each other, and forced-choice performance was compared against chance (both *t*-test analyses). Due to limited open-ended variability, subsequent analyses were conducted with forced-choice data only.

We next examined whether there were differences in self-derivation performance and error variance between grade levels. Levene's test for equality of error variance was significant, indicating a grade-level difference in error variance. Accordingly, we conducted separate mixed-factor ANOVAs for Grade 3 and Grade 4 with education model (traditional or bilingual) as the between-participants factor and condition (control or manipulated) as the within-participant factor. The dependent measure was forced-choice accuracy score, calculated based on the number correct out of the total number of trials per condition.

We next examined whether there were grade-level differences in verbal comprehension using another mixed-factor ANOVA with grade (3 or 4) as a between-participants factor and language (English or Spanish) as the within-participant factor.

Finally, we examined individual differences in self-derivation performance. We conducted regression analyses to examine the unique predictive power of English verbal comprehension and Spanish verbal comprehension as well as school experience, indicated by grade level. Age was not included in these models because grade was available as a better proxy for experience in the academic context, which was the variable of interest (for discussion, see Morrison, Griffith, & Alberts, 1997). Just as in Study 1, there were two models. In Model 1, we examined total performance for all participants across education models and language conditions (and grades) to test the hypothesis that English verbal comprehension would predict performance in English. In Model 2, we specifically tested the hypothesis that both English and Spanish verbal comprehension would predict performance in the cross-language condition by examining performance of the cross-language stimuli (only the bilingual education participants in both grades contributed data).

## Results

The results are reported in two sections. We first examined children's self-derivation of new factual knowledge through integration across grade, education model, and conditions. Second, we assessed the role of verbal comprehension as a predictor of successful self-derivation through integration. All analyses were conducted using the SPSS Statistics package (Version 24). All statistical tests were two-tailed. Bonferroni corrections were made when appropriate for multiple comparisons.

*Self-derivation through integration*

Self-derivation of new factual knowledge through integration of separate yet related episodes of new learning is reported by grade, education model, and condition in Table 4 (Panel B). Performance is reported as percentage correct because the number of trials differed between conditions (7 or 8 based on counterbalancing procedures). As is apparent from the table, open-ended questioning revealed little evidence of self-derivation of new factual knowledge through integration. Performance averaged under 10% for third graders and under 20% for fourth graders. Again, errors of commission were too few to analyze. Children had significantly higher performance in forced-choice questioning,

$t(99) = -23.23$, $p < .001$. Performance in forced-choice questioning was significantly above chance across grades and conditions ($ts > 3.49$, $ps < .002$), with the exception of the third-grade cross-language condition ($t = 0.78$, $p = .44$) (chance = 33%), and resembled performance of adults' forced-choice performance in similar procedures (e.g., Bauer & Jackson, 2015). In subsequent analyses, we focused on forced-choice performance only due to the limited range exhibited in open-ended testing.

As noted above, Levene's test for equality of error variance was significant in a mixed-factor ANOVA, indicating that error variance differed by grade level, $F(3, 89) = 3.66$, $p = .02$. Thus, we conducted separate mixed-factor ANOVAs for Grade 3 and Grade 4.

In the Grade 3 analyses, education model was not significant, $F(1, 41) = 0.77$, $p = .38$. There was a main effect of condition such that children's scores were higher in the control condition than in the manipulated condition, $F(1, 41) = 7.59$, $p = .009$, $\eta^2 = .16$. The main effect was qualified by a significant education model by condition interaction, $F(1, 41) = 5.56$, $p = .02$, $\eta^2 = .12$. Examination of the interaction revealed higher levels of performance in the control condition than in the manipulated condition only for children in the bilingual education program. Thus, children in the traditional education program did not show performance differences in the control condition (English–English, same speaker) compared with the manipulated condition (English–English, different speakers). However, children in the bilingual education model did show differences between the control condition (English–English, same speaker) and manipulated condition (English–Spanish, different speakers). Moreover, examination of the interaction by condition revealed that within the manipulated condition children in the bilingual education program (English–Spanish, different speakers) had lower level of performance compared with children in the traditional education program (English–English, different speakers). The education program groups did not differ in the control condition (English–English, same speaker).

In the Grade 4 analyses, there were no significant main effects or interaction, $Fs(1, 48) \leq .01$, $ps \geq .91$. Thus, whereas performance differed between the control and manipulated conditions for Grade 3, it did not for Grade 4.

*Verbal comprehension*

The difference in performance between grade levels on the integration fact questions suggests that fourth-grade children had higher Spanish proficiency relative to third-grade children. However, examination of the means for children in the bilingual education program counters this interpretation because mean Spanish performance for third graders was nominally higher than that for fourth graders. Analysis of verbal comprehension within the bilingual education program in a grade (3 or 4) by language (English or Spanish) mixed-factor ANOVA revealed no effect of grade. Across grades, children had higher scores on English verbal comprehension compared with Spanish verbal comprehension, $F(1, 42) = 4.59$, $p = .04$, $\eta^2 = .10$; the interaction with grade was not significant.

To examine possible relations between self-derivation of new factual knowledge through integration and verbal comprehension, we tested regression models. Results are reported in Table 5 (Panel B). The first two models predicted total performance across education programs and conditions (all participants), first with grade as a predictor and second adding English and Spanish verbal comprehension. The first model was not significant, indicating that grade level was not a significant predictor of total self-derivation performance. The second model was significant, replicating the results of Study 1 such that across programs English verbal comprehension was the only significant predictor of total self-derivation performance. When we repeated the analysis with only cross-language self-derivation performance (one stem fact through Spanish and one stem fact through English, only participants from the bilingual program), both models were significant. Grade was a significant predictor of cross-language self-derivation. When verbal comprehension was added to the model, grade and Spanish verbal comprehension, but not English verbal comprehension, were significant predictors of cross-language self-derivation. Thus, English verbal comprehension predicted performance when the majority of facts were presented through English, but both grade and Spanish verbal comprehension were significant predictors when half of the facts were presented through Spanish.

*Discussion*

The results of Study 2 were consistent with our predictions that the low surface similarity of self-derivation through cross-language integration could pose a challenge. The results indicate that this is especially difficult for younger children with less experience in a bilingual learning environment. Specifically, for third-grade children in the bilingual program, the requirement to integrate separate lessons across languages was associated with lower levels of performance relative to both control conditions: (a) their own English–English performance and (b) English–English performance of children in the traditional education program. In contrast, fourth graders' cross-language performance did not differ from either English–English control condition. There were no differences between the control (same speaker) and manipulated (different speakers) conditions for children in the traditional education model.

The most obvious explanation for the expected pattern of performance is in terms of Spanish-language proficiency. The difference in the stimuli between the control and manipulated conditions for children in the bilingual program was the addition of Spanish stem facts. Thus, children may have had trouble in comprehending the Spanish facts, explaining the lower performance in the cross-language condition compared with the same-language conditions. This is supported by the lack of difference between conditions in fourth grade when children had completed an extra year of Spanish instruction and, thus, may have gained additional proficiency in the language. Yet, this explanation is undermined by the finding that there were not differences in either English or Spanish verbal comprehension between the children in Grades 3 and 4 as measured by the WMLS–RNU raw scores. The absence of differences implies that (a) there is variability in vocabulary that is independent of grade level and (b) it is unlikely that the grade groups differed in the comprehension of items presented at test. Although the children are likely continuing to improve their Spanish proficiency after third grade in ways not captured by the WMLS–RNU measure, these results support that the differences in performance are not solely the result of differences in Spanish proficiency.

We suggest that the pattern of performance across grades may also be attributed to differences in the amount of practice and experience the children have had in integrating across low surface similarity conditions (i.e., languages, teachers, classrooms, materials). In the school system where testing took place, children enter the bilingual education program in kindergarten. By the time they are in Grade 4, children in the bilingual education program have had 5 years of experience in learning through two languages. Children in Grade 3 have had 1 less year of experience. The day-in-and-day-out requirement to integrate across languages, teachers, classrooms, materials, and cultural representations, as well as recognizing conceptual representations across these low surface similarity conditions, likely contributes to a domain-general ability that is more developed in children with more experience. The data support this; grade predicted significant unique variance in the cross-language regression model even with the inclusion of verbal comprehension. Thus, rather than a vocabulary-specific effect, we propose an additional domain-general effect as an explanation for the pattern of findings.

## General discussion

In the current research, we examined self-derivation of new factual knowledge through integration of separate yet related episodes of new learning as a function of differences in the surface similarity of presentation created by the language (or languages) in which the material was presented. In two experiments, children were presented with pairs of facts that could be integrated to support the self-derivation of new knowledge. Across both studies, there was a control condition in which all information was presented in English. Across both studies, the control conditions did not differ by education program. Children also participated in a surface similarity manipulation. For children in the traditional education program (English instruction), they heard the related facts from two different speakers and performance did not differ from the control condition with a single speaker. Children in the bilingual program (Spanish and English instruction) participated in a cross-language surface similarity manipulation. In Study 1, children were provided with contextual support in the context of four related story pairs, including unique characters, plot, and setting for each story, and there were no differences in performance (cross-language vs. single-language controls). In Study 2, children in Grades 3 and 4 were presented with 8 single-sentence fact pairs in the manipulation condition without the contextual support of text passages. Here, an examination of forced-choice performance revealed a "cost" to the low surface similarity conditions of cross-language self-derivation through integration performance for third graders but not for fourth graders.

The results ran counter to predictions in two ways. First, among second-grade and fourth-grade children, there was not an effect of the cross-language manipulation. In Study 1, children in the early stages of second-language learning did not show a cost in self-derivation through integration across languages. This is inconsistent with the expectation that children would struggle to recognize that related facts could be integrated when surface similarity was reduced by presenting the facts through different languages. In previous research, children were challenged by a change in character between two related stories (Bauer et al., 2012), and a change in language, especially with limited proficiency in one of the languages of presentation, was expected to be difficult. Yet, in Study 2, when the rich context was removed through use of a single-sentence format (rather than story passages), third-grade children were challenged by the change in language between related facts. The single-sentence format was generally difficult for children in both grades, as evidenced by the low open-ended performance. Together, the findings could be taken to suggest that the context provided in Study 1 was sufficient to support cross-language integration even for young children who typically depend on high surface similarity to recognize the relevance of related facts and with limited proficiency in the two languages of presentation.

Second, the expectation that language proficiency would account for the pattern of results was not fully supported in Study 1. Specifically, English verbal comprehension was a significant predictor of performance for both the total performance model and cross-language model. In addition, Spanish verbal comprehension did not predict performance in either model. If language proficiency alone were carrying performance, we would expect a contribution of Spanish verbal comprehension in the cross-language condition where half of the stem facts required to self-derive through integration were presented through Spanish. This underscores our interpretation of the lack of differences as being due to the benefits of contextual support.

Similarly, in Study 2, language proficiency alone did not explain the different pattern of results across grade levels. Children in fourth grade did not have higher Spanish verbal comprehension than third graders as measured by the WMLS–RNU. Thus, the expectation that higher rates of cross-language self-derivation through integration in fourth grade (above chance) compared with third grade (at chance) could be attributed solely to gains in language proficiency was not corroborated. Verbal comprehension is one supporting component, as evidenced by the significant and unique contribution of Spanish verbal comprehension in predicting cross-language performance (31% of the variance). However, the results also revealed a significant contribution of grade in the cross-language models (37% of the variance in the final model). In addition to growth in verbal comprehension, children gain experience in a bilingual environment between third grade and fourth grade, with grade as a proxy for time spent in this learning environment. It could be interpreted that children with more experience in bilingual education were better able to navigate cross-language self-derivation through integration even in a paradigm with limited contextual support. This interpretation is similar to the established finding that experience in an academic context (years in school) is responsible for academic gains made across the school year rather than children's age (Morrison et al., 1997). Just as children learn how to "do school," the fourth-grade children in Study 2 may have learned how to "do cross-language."

The interpretation that time in bilingual education supports cross-language integration is further supported by the pattern of results across the studies and conditions. In Study 1, English verbal comprehension predicted self-derivation through integration performance. In Study 2, this finding was replicated in the model predicting total performance, with English verbal comprehension and not grade or Spanish verbal comprehension predicting significant unique variance. However, in the cross-language conditions, we saw a change in the pattern of results. Unlike the other models, in the Study 2 cross-language model grade and Spanish verbal comprehension were significant predictors and English verbal comprehension dropped out as a significant and unique predictor. This pattern further supported our interpretation that experience in a bilingual learning environment, along with verbal comprehension, predicts cross-language self-derivation performance.

Based on the observed pattern of relations, an interesting future direction could be to examine self-derivation performance as it is relates to relative between-language proficiencies. Consider that children could have high levels of proficiency in both languages, low levels of proficiency in both languages, or high proficiency in one language and low proficiency in the other. We would expect a child with high proficiency in both languages to perform well, whereas a child low in verbal comprehension in both Spanish and English would struggle with self-derivation. Perhaps the more interesting case is a child with high proficiency in one language and low proficiency in the other. In this situation, the question is whether high proficiency in only one language would be sufficient for success. Moreover, how would the language of presentation interact with this pattern of proficiency? Unfortunately, we do not have a sufficient number of participants to divide the sample into quartiles. Yet, the results from Study 2, in which we found that each language was a unique predictor when the majority of stimuli were presented in that language but not when the majority of stimuli were presented in the other language, suggest that the specific language of stimulus presentation is important to self-derivation performance.

The results should be interpreted with some caution. In Study 1, the number of trials that could be tested with the young participants was limited, resulting in a restricted range for the variable of interest. This was rectified in Study 2, but the necessary change in protocol between Study 1 and Study 2, in addition to the cross-sectional nature of the data, necessitates caution in interpreting differences in performance between the studies.

In conclusion, under most circumstances, children were able to integrate and extend knowledge through self-derivation when facts were provided through different languages, indicating that cross-language fact presentation is not necessarily a barrier to self-derivation through integration. The results indicate that cross-language self-derivation through integration is a multifaceted challenge that, along with verbal comprehension, can potentially be mitigated by contextual support and grade-level-related experiences. The findings are important to our understanding of the role of surface similarity in productive processes generally and in self-derivation through integration specifically.

## Acknowledgments

This work was supported by grants from the National Institute of Child Health and Human Development (HD067359), the National Science Foundation (BCS1528091), and the Institute of Education Sciences (R305A150492) to Patricia J. Bauer and by Emory College of Arts and Sciences. The authors thank the participating school system and the families who took part in the research.

## Appendix A. Participating dual-language program

The questions of interest required a sample of children with experience in two languages. Such a sample was available in a school system that offers an English–Spanish bilingual education option. Children whose parents are interested in bilingual education enter their children into a lottery at the time of kindergarten registration. Children are selected into the bilingual program pseudorandomly, maintaining a balance of home language, sex, and racial/ethnic distribution. Regardless of language background (English or Spanish as the home language), children who are not selected into the lottery attend traditional single-language (English) classrooms side by side with the bilingual classes. Bilingual education placement is a stable assignment from kindergarten to the completion of fifth grade, with few children leaving the program (all children in the current study had been in the program since kindergarten).

The participating bilingual education program is a two-way dual-immersion 50/50 program. Classes are composed of equal parts English- and Spanish-speaking children, and content is presented through English and Spanish, each 50% of the time. In this program, children alternate days of instruction between two "language worlds." They spend a day immersed in English with all subjects taught through English, followed by a day immersed in Spanish with all subjects taught through Spanish, and so forth. Importantly for current purposes, lessons are not repeated across languages; rather, the content builds with each lesson such that children must integrate and extend between two languages to successfully construct a knowledge base. Children enrolled in the program begin primarily as Spanish or English monolingual speakers, according to parent report, but acquire the partner language as they progress through their elementary school career. Rather than being taught a language, children are taught grade-level content *through* a language. Although maintenance of separate language environments is a goal, in practice partitioning is not perfectly maintained, resulting in a bilingual learning environment.

## References


Anolli, L., Antonietti, A., Crisafulli, L., & Cantoia, M. (2001). Accessing source information in analogical problem-solving. *Quarterly Journal of Experimental Psychology A, 54*, 237–261.

Bauer, P. J. (2012). Facilitating learning and memory in infants and young children: Mechanisms and methods. In S. L. Odom, E. P. Pungello, & N. Gardner-Neblett (Eds.), *Infants, toddlers, and families in poverty: Research implications for early child care* (pp. 49–74). New York: Guilford.

Bauer, P. J., Blue, S. N., Xu, A., & Esposito, A. G. (2016). Productive extension of semantic memory in school-aged children: Relations with reading comprehension and deployment of cognitive resources. *Developmental Psychology, 52*, 1024–1037.

Bauer, P. J., Dugan, J. A., Varga, N. L., & Riggins, T. (2019). Relations between neural structures and children's self-derivation of new knowledge through memory integration. *Developmental Cognitive Neuroscience, 36*, 100611.

Bauer, P. J., Esposito, A. G., & Daly, J. J. (2019). Recall of episodic and semantic details: Self-derivation through integration after a delay. *Learning and Instruction*. Submitted for publication.

Bauer, P. J., & Jackson, F. L. (2015). Semantic elaboration: ERPs reveal rapid transition from novel to known. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 271–282.

Bauer, P. J., King, J. E., Larkina, M., Varga, N. L., & White, E. A. (2012). Characters and clues: Factors affecting children's extension of knowledge through integration of separate episodes. *Journal of Experimental Child Psychology, 111*, 681–694.

Bauer, P. J., & Larkina, M. (2017). Realizing relevance: The influence of domain-specific information on generation of new knowledge through integration in 4- to 8-year-old children. *Child Development, 88*, 12584.

Bauer, P. J., & San Souci, P. (2010). Going beyond the facts: Young children extend knowledge by integrating episodes. *Journal of Experimental Child Psychology, 107*, 452–465.

Brown, A. L. (1982). *Learning, remembering, and understanding* Technical report no. 244. Center for the Study of Reading, University of Illinois at Urbana-Champaign.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology, 20*, 493–523.

Brown, A. L., Kane, M. J., & Echols, C. H. (1986). Young children's mental models determine analogical transfer across problems with a common goal structure. *Cognitive Development, 1*, 103–121.

Center for Applied Linguistics. (n.d.). Dual language program directory. Washington, DC: Author. Available at: http://webapp.cal.org/DualLanguage/DLListing.aspx.

Center for Public Education (2015). *Learning to read, reading to learn: Why third-grade is a pivotal year for mastering literacy*. Alexandria, VA: Center for Public Education.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49*, 222–251.

Cummins, J. (2000). *Language, power, and pedagogy: Bilingual children in the crossfire (Bilingual Education & Bilingualism, Vol. 23)*. Bristol, UK: Multilingual Matters.

Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist, 47*, 153–176.

Esposito, A. G., & Bauer, P. J. (2017). Going beyond the lesson: Self-generating new factual knowledge in the classroom. *Journal of Experimental Child Psychology, 153*, 110–125. https://doi.org/10.1016/j.jecp.2016.09.003.

Esposito, A. G., & Bauer, P. J. (2018). Building a knowledge base: Predicting self-derivation through integration in 6- to 10-year-olds. *Journal of Experimental Child Psychology, 176*, 55–72.

Gentner, D. (1977). Children's performance on a spatial analogies task. *Child Development, 48*, 1034–1039.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development, 59*, 47–59.

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on language and thought: Interrelations in development* (pp. 225–277). Cambridge, UK: Cambridge University Press.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science, 10*, 227–300.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*, 306–355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1–38.

Goswami, U. (1992). *Analogical reasoning in children*. New York: Psychology Press.

Goswami, U. (2011). Inductive and deductive reasoning. In U. Goswami (Ed.), *Childhood cognitive development* (pp. 399–419). Oxford, UK: Wiley-Blackwell.

Hoff, E. (2013). *Language development*. Boston: Cengage Learning.

Kids Count Data Center (2018). *The number of bilingual kids in America continues to rise*. Baltimore, MD: Annie E. Casey Foundation https://datacenter.kidscount.org/updates/show/184-the-number-of-bilingual-kids-in-america-continues-to-rise.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67*, 2797–2822.
Mix, K. S. (1999). Similarity and numerical equivalence: Appearances count. *Cognitive Development, 14*, 269–297.
Morrison, F. J., Griffith, E. M., & Alberts, D. M. (1997). Nature–nurture in the classroom: Entrance age, school readiness, and learning in children. *Developmental Psychology, 33*, 254–262.
Perret, P. (2015). Children's inductive reasoning: Developmental and educational perspectives. *Journal of Cognitive Education and Psychology, 14*, 389–408.
Poarch, G. J., Van Hell, J. G., & Kroll, J. F. (2015). Accessing word meaning in beginning second language learners: Lexical or conceptual mediation? *Bilingualism: Language and Cognition, 18*, 357–371.
Siegler, R. S. (1989). Mechanisms of cognitive development. *Annual Review of Psychology, 40*, 353–379.
Varga, N. L., & Bauer, P. J. (2013). Effects of delays on 6-year-old children's self-derivation and retention of knowledge through integration. *Journal of Experimental Child Psychology, 115*, 326–341.
Varga, N. L., Esposito, A. G., & Bauer, P. J. (2019). Cognitive correlates of memory integration across development: Explaining variability in an educationally relevant phenomenon. *Journal of Experimental Psychology: General, 148*, 739–762.
Varga, N. L., Stewart, R. A., & Bauer, P. J. (2016). Integrating across episodes: Investigating the long-term accessibility of self-generated knowledge in 4-year-old children. *Journal of Experimental Child Psychology, 145*, 48–63.
Verhoeven, L. T. (1994). Transfer in bilingual development: The linguistic interdependence hypothesis revisited. *Language Learning, 44*, 381–415.
Verhoeven, L. (2007). Early bilingualism, language transfer, and phonological awareness. *Applied Psycholinguistics, 28*, 425–439.
Winner, E., Rosenstiel, A. K., & Gardner, H. (1976). The development of metaphoric understanding. *Developmental Psychology, 12*, 289–297.