LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Input Variability in Foreign Language Verb–Argument Construction Learning: A Primary School Intervention

Johannes Schulz,[a,b] Victoria Murphy,[a] and Elizabeth Wonnacott[a]

[a]University of Oxford [b]University of Education Freiburg

Abstract: *Input variability* (i.e., varied experience with different grammar exemplars; "[talk/think/rant/wonder] about [god/bicycles/dogs]") can improve generalization (i.e., "[Verb] about [Noun]") and enhance learning. Controlled experiments show that input variability benefits children's generalization and extension of linguistic information from input structures to novel contexts (Wonnacott et al., 2012). Such findings drove our investigation into extending input variability benefits to classrooms. Following a usage-based constructionist approach, we present a two-week quasi-experimental teaching intervention with two British Year 2 classes learning German (6-to-7-year-olds; 20 students per class). The intervention featured a variability and no-variability condition,

CRediT statement – **Johannes Schulz**: conceptualization, investigation, methodology, formal analysis, data curation, visualization, resources, project administration, writing – original draft, writing – review & editing. **Victoria Murphy**: conceptualization, methodology, supervision, funding acquisition, writing – review & editing. **Elizabeth Wonnacott**: conceptualization, methodology, supervision, funding acquisition, writing – review & editing.

A one-page Accessible Summary of this article in nontechnical language is freely available in the Supporting Information online and at https://oasis-database.org.

The authors declare no competing interests.

We would like to thank the host school, the host teacher, and all participating children for their invaluable contribution to this study. We also thank the handling editor, Dr. Murakami, and the five anonymous reviewers for their helpful comments on earlier versions of the manuscript.

Correspondence concerning this article should be addressed to Johannes Schulz, Institute of Psychology, University of Education Freiburg, Kunzenweg 21, Freiburg, 79117, Germany. Email: johannes.schulz@ph-freiburg.de

The handling editor for this article was Akira Murakami.

DOI: 10.1111/lang.70051

using 16 German "approach" event verb–argument constructions ("To the X [verb] the Y"), featuring four separate or just one verb in the verb slot, respectively. Posttests indicated that children exposed to variability generalized better to novel verbs. The findings corroborate controlled experiments. Even in "noisy" classrooms, input variability can benefit students' construction generalization and extension to novel verbs.

Keywords input variability; primary school; German as a foreign language; intervention; communicative agency

Introduction

Despite a worldwide uptake in primary school foreign language (FL) instruction since the 1980s (Enever, 2011), FL communication skills by the end of primary school are often considered rudimental (Hempel, 2016), and many students are reported to be stuck on reproductive and imitative plateaus (Engel et al., 2009). This development contrasts with curricular expectations that prioritize a high degree of communicative agency for young FL learners (Department for Education, 2013), where *communicative agency* refers to the learners' ability to select and organize language so that it corresponds to their individual communicative intent (cf. Doyé & Lüttge, 1977).

Usage-based approaches posit that FL acquisition is a largely implicit and incidental input-driven process (Madlener-Charpentier & Behrens, 2022). In the context of meaningful communicative interactions, learners gradually reconstruct the target language's constructional repertoire across encounters with individual instantiations of *constructions* (e.g., Ellis, 2009, 2017; Ellis & Wulff, 2020; Madlener-Charpentier & Behrens, 2022), so-called form–function pairings, such as ditransitive constructions (Goldberg, 1995; Jackendoff, 2002). In input-limited primary FL contexts (i.e., in the early stages of instructed learning), learners initially rely on a small set of relatively fixed multiword patterns to establish basic, situation-dependent, and mostly imitative communicative agency. For example, target constructions at the phrasal level that are ubiquitous in primary classrooms include routines (*How are you?*) or patterns (*My favourite _ is _*) (Kostka, 2020). For learners' linguistic repertoires to become more productive, they must move beyond this initial phase of situation-dependent, largely imitative communication and develop a more situation-independent communicative agency. This allows learners to express communicative intentions independent of specific situations. This development of increasingly situation-independent communicative agency, in addition to the imitative use of fixed multiword units, depends on the degree of productivity of a learner's linguistic repertoire (Kostka, 2020). Specifically, the degree of productivity relates to the degree of abstractedness of a learner's systematic mental network of productive constructional abstractions, such as verb–argument constructions (VACs). The acquired constructional abstractions equip the learner with communicative flexibility. However, the development of such abstract linguistic representations is contingent upon a learner's ability to induce systematicities across individual occurrences of specific linguistic patterns in a variety of meaningful interactions (Goldberg & Casenhiser, 2008). A typical induction process where linguistic units are encountered in a variety of contexts could develop from concrete units (e.g., *I don't know*) via semifixed formulas (e.g., *I don't + Verb*), to abstract representations like syntactic schemas, such as "Subject Verb (Object)" (Bannard et al., 2009; Ibbotson, 2013). In the end, learners acquire an at least partially abstract generalization (Goldberg, 2009). Thus, one tenet of a usage-based, constructionist approach to language learning, and the process of establishing communicative agency that goes beyond imitation, is the ability to arrive at schematizations and generalize productive, partially abstract patterns (Goldberg, 2009). Importantly, this precise notion is reflected in current primary school FL curricula worldwide. On the one hand, phrasal patterns from the teaching input are considered crucial initial means in establishing rudimental communication (MSB NRW, 2012). On the other hand, they are highlighted as key tools for helping students generalize language experiences and form hypotheses about pattern formation (MSB NRW, 2012). These patterns also enable students to decouple language from specific situations, fostering situation-independent knowledge that enhances their communicative abilities (KM-BW, 2004). Given the discrepancy between expected learning outcomes (i.e., communicative agency beyond imitation) and actual learning outcomes (i.e., "imitative plateau"), one core question for educators and researchers is how to manipulate the teaching input to minimize this gap. A better understanding of the role of input variability in young FL learners, therefore, is an important first step.

Background

Under input-driven accounts of learning, such as usage-based language learning, we would expect generalization to be closely tied to the structure of the input which learners encounter (Pagonis, 2022). Specifically, we expect generalization to be impacted by *input variability*, that is, the variation in the different exemplars of the construction encountered. When overall token frequency is held constant, low type frequency input provides more repetitions per token, strengthening entrenchment and surface similarity (Ellis &

Ferreira-Junior, 2009). This supports storage and alignment processes that help learners detect patterns (i.e., pattern detection) (Ellis & Ferreira-Junior, 2009). However, limited type variability may constrain abstraction and generalization beyond the input. In contrast, high type frequency may promote schematization and productive pattern extension, as learners observe broader variability and can infer what information is relevant versus irrelevant for linguistic generalization (e.g., Madlener-Charpentier, 2015). This role of (high) type frequency in driving linguistic generalization (Bybee, 1995, 2008)—is predicted by accounts in which language learning (like other types of learning) is underpinned by a process of uncertainty reduction which then leads to the discrimination of reliable cues through the elimination of the alternatives (Ramscar, 2021; Ramscar et al., 2010).

The benefits of (high) type frequency have been reported in a range of cognitive science domains (Raviv et al., 2022). Specifically, with linguistic generalization, Gómez (2002) and Gómez and Maye (2005) demonstrated that infants and adults were able to better detect nonadjacent dependencies of the form A-X-D (e.g., pel-wadim-rud), B-X-E (e.g., vot-wadim-jic), and C-X-F (e.g., dak-wadim-tood), where X denotes any two-syllabic word, when X was drawn from a larger sample of words, a situation which increased variability and differentiated the irrelevant X-item from the key invariant, nonadjacent dependencies. In the domain of word learning, Twomey et al. (2014) found that 2-year-olds presented with multiple exemplars of a category demonstrated better retention of label–object associations than children who encountered only the same exemplar repeatedly. The authors argued that children who encountered multiple exemplars retained words at greater rates because the act of comparing exemplars across trials enabled children to identify category-relevant features while downplaying category-irrelevant features. Viviani et al. (2026) explored whether 7- to 8-year-olds could learn postpositions meaning “above” and “below” in a novel language (Japanese) by playing a game where they moved pictures around a grid (e.g., moving a picture of an ice cream above a picture of a banana in response to an instruction in Japanese). They found better generalization (i.e., the children could better respond to the instructions involving the same postpositions but different nouns) if children were trained using more varied instruction sentences than if they were trained on a small set of repeated instructions.

Turning to VACs, Casenhiser and Goldberg (2005) exposed 5- to 7-year-olds to an argument–structure construction mapping a novel meaning (*appearance*) onto a novel form (SOV word order). Results showed that, compared to an equal distribution of novel verbs, generalization of the abstract construction onto new vocabulary was stronger when the input was skewed toward one particular nonsense verb. These findings aligned with similar studies on adult learners (Goldberg et al., 2004). A related classroom-based study on skew effects was done by Year and Gordon (2009), who examined Korean learners of English (aged 12–13) in their acquisition of ditransitive constructions, featuring five verb types in total.[1] They found no clear skew benefit but also no evidence against it. Instead, the data patterned in the direction that learners benefited more from balanced variable input. This stands in contrast to findings from more controlled settings and warrants further empirical investigation. Importantly, however, this study does not provide evidence for a general advantage of more varied input over invariant input in primary instructional settings. Rather, Year and Gordon (2009) manipulated the distribution of tokens across verb types without varying verb type frequency itself.

Notably, Casenhiser and Goldberg (2005) did not measure children's ability to extend their constructional knowledge to novel contexts in production. Testing only required children to react to test sentences. Wonnacott et al. (2012) addressed this by extending the outcome measures to capture productive generalization. In addition, they manipulated verb type frequency across conditions, though they did not vary the within-condition frequency distribution (e.g., skewed input). Specifically, Wonnacott et al.'s (2012) artificial language study investigated whether a targeted variability manipulation of the input would impact young learners' (mean age = 5;2) ability to extend acquired generalizations to novel vocabulary. They used a semiartificial language with a novel approach–event construction (VerbNoun$_1$Noun$_2$; "Noun$_2$ approaches Noun$_1$ in the manner denoted by Verb"), such as *chadding rabbit gorilla* where *chadding* means "(to) hop on one's head." During exposure, an experimenter read out input sentences while enacting the events with toy animals (e.g., saying "Chadding rabbit gorilla" and bouncing the gorilla on its head while moving it towards the rabbit). Half of the children encountered sentences where the same verb occurred in the verb slot on each trial (the no-variability group), while the other group encountered different (novel) verbs in each trial (the variability group). The variability group encountered four different verbs in total. After three days of training, children were tested in act-out, production, and forced-choice tasks. Responses to the former two tasks were coded as either "approach incorrect" (approach correct but reverse linking), "approach correct" (approach plus linking rule correct), or "other."

The last category included instances where children carried out actions that did not fit a ditransitive pattern or did not move the animals. In both the act-out and production tests, children in the variability condition were more likely than those in the no-variability condition to produce either an "approach correct" or "approach incorrect" response, rather than an "other" response. The authors took this as evidence to mean that children were more able to generalize the global meaning of the construction following variability training. Using stricter coding, which looked at the likelihood of producing an "approach correct" response (rather than either an "approach incorrect" or "other" response), there was again a variability advantage for unfamiliar verbs in the act-out task, but not in the production task. In the forced-choice tasks, where Wonnacott et al. (2012) tested linking rules only, no evidence for a difference between conditions was found. Based on their findings, the authors concluded that verb type frequency generally facilitates generalization, at least for learning the global semantics of the construction (i.e., learning that the VNN pattern maps to a transitive approach event). They argued that learning the linking rule was more challenging given the brief exposure and given the children's strong bias (from English) to assume that the first mentioned noun would be the agent.

There is thus evidence from language learning studies, and from other domains (Raviv et al., 2022), that suggests a role for variability in driving generalization. Yet, note that variability effects have not been consistently observed across studies. For instance, Brown et al. (2021) examined an artificial language where particle words, somewhat analogous to gender markers or classifiers in natural languages, were linked to semantic features. While both 7-year-olds and adults demonstrated some ability to learn these semantic associations, exposure to a greater variety of input examples, where particles were paired with a larger set of nouns (four nouns vs. eight nouns), did not lead to enhanced generalization. The authors suggested that this was due to the highly explicit nature of the learning task (as confirmed by posttest questionnaires), whereas the variability effect is generally driven by implicit learning mechanisms.

These caveats aside, there seems to be evidence that, at least in some circumstances, increased type frequency can benefit generalization. Increased type frequency might thus be a promising variable to manipulate in the input of primary school FL classrooms. In fact, many have considered targeted manipulations of type and token frequencies in the input to have important implications for FL teaching. For instance, Robinson and Ellis (2008, p. 509) inquired, "How do type and token frequency interact in learning SL [Second

Language] constructions? (…) Does increasing type variation in the verb slot lead to greater abstraction and generalizability of argument structure constructions?" Considering the worldwide rise in primary school FL instruction, these questions are crucial for FL research. However, relevant primary school evidence in this regard is scarce, as demonstrated in a systematic review by Schulz et al. (2023). Of the studies discussed above, all of which employed artificial language input, only Wonnacott et al.'s study (2012) was conducted fully in a school, but with the experimenter training each child individually. Viviani et al.'s study (2026) was partially conducted in schools, with children tested in whole class groups, but the training and testing were computerized. In fact, concerns have been raised whether findings from cognitive psychology regarding distribution effects generalize to "the domain of SLA, that is, natural languages and the contexts in which they are acquired" (Robinson, 2005, p. 262). In such noisy settings, learners must (a) acquire a natural language and (b) navigate a plethora of potentially distracting linguistic (e.g., classmates talking) and nonlinguistic stimuli (e.g., movement in the room) in addition to the target patterns. Specifically, the everyday distractions of a primary classroom can disrupt students' attention to linguistic stimuli, thereby impacting their learning process. In contrast, in Wonnacott et al.'s study (2012)—and others similar—the children were exposed to stimuli individually, away from their classmates, making the setting more tightly controlled. In addition, there was little attempt to contextualize the language, which may have led participants to (attempt to) recognize the study's purpose. Consistent with usage-based theories on the importance of meaningful context, a natural input setting may, in contrast, promote more implicit, incidental learning rather than the potentially explicit processes which are often inadvertently encouraged by the contrived nature of laboratory tasks (Madlener-Charpentier, 2018). Finally, in artificial language tasks, stimuli are often tailored to the participants' L1 (e.g., Wonnacott et al., 2012), whereas natural language input may introduce an extra layer of difficulty when inferring linguistic patterns from the input (e.g., German phonological characteristics).

In light of these points, it is important to establish whether findings from controlled, artificial language experiments are relevant in FL primary contexts. In the current study, we focused on the verb-type frequency effect reported by Wonnacott et al. (2012). This is because they used outcome measures that captured participants' active generalization of abstracted patterns to novel verbs and did so with children of an age range relevant to primary contexts. The inclusion of a 1-verb condition, that is, a condition that featured only one verb in the target construction's verb slot, served as a baseline against which effects

Table 1 Descriptive statistics of background tests

	WASI			PVST			Language Magician		
condition	*M*	*SD*	*MD*	*M*	*SD*	*MD*	*M*	*SD*	MD
variability	14	6.1	16	4,142	537	4,094	56	8.8	56
no variability	15	4.3	15	4,329	355	4,375	53	6.6	52

Note. Picture Vocabulary Size Test (PVST) expected score for this age group is approximately 4,000 (Anthony & Nation, 2021). The expected age equivalent raw scores for the Wechsler Abbreviated Scale of Intelligence (WASI) Matrix Reasoning subtest range from 8 points for children aged 6;2 years to 12 points for children aged 7;6 years (Wechsler, 2011).[3] Expected scores for the Language Magician are not available.

of increased type frequency could be evaluated, particularly in a noisy primary FL classroom context. In this sense, the study provides a proof of concept that frequency-based learning effects can emerge under such conditions. Thus, our research question (RQ) is:

> To what extent does input variability support instructed primary FL learners' ability to generalize the input's form–function correspondence and extend this knowledge to novel contexts, thereby increasing their communicative agency?

Methodology

Participants and Design

This classroom intervention experiment was conducted in a British primary school with two intact classes.[2] The participants were 40 Year 2 students (6-to-7-year-olds) who had had German FL instruction since Year 1 for one lesson each day (lasting 30 minutes) with a teacher who spoke German natively. A quasi-experimental design was used whereby each of two intact classes was assigned to either a no-variability ($n = 1$ class, $n = 20$ students) or variability ($n = 1$ class, $n = 20$ students) input condition. A battery of background tests was conducted to determine whether the classes were roughly matched on language ability (in English and German) and tests of cognition. The results from these background tests are shown in Table 1. Both groups undertook the intervention followed by a series of three posttests (act-out, production, forced choice) which formed the datasets for analyses. Note that this quasi-experimental design is somewhat limited given that the sample size at the class level was, effectively, $n = 2$, meaning that there was no controlling for clustering of the data at class level. Therefore, we consider this experiment a proof of concept.

Materials and Stimuli

Background Tests

Following previous FL teaching interventions in primary schools (Hopp & Thoma, 2021), baseline assessments included a cognitive abilities test (the Matrix Reasoning subtest of the Wechsler Abbreviated Scale of Intelligence for Children [WASI]; Wechsler, 1992) and German FL (the Language Magician) and English (Picture Vocabulary Size Test [PVST]; Nation & Anthony, 2016) language tests. The Language Magician is a language assessment game for primary students that was developed by the European Union and research institutes across Europe. It provides teachers with a child-appropriate tool to measure FL ability in various languages, including German. The assessment taps into reading comprehension, listening comprehension, and writing skills. All of the background tests were computerized tasks, except for the WASI, which children did on paper in class, working simultaneously. This approach did not fully comply with test instructions. Details on the WASI procedure are available in Appendix S1 in the Supporting Information online.

We did not include a background test for target construction knowledge, as it was reasonably assumed to be unknown and the participants' low German proficiency made an additional test impractical.

Teaching Intervention

Similar to Wonnacott et al.'s study (2012), we introduced variability in a VAC's verb slot to investigate whether input variability in this slot impacts students' learning of global construction semantics and linking rules. Also similar to Wonnacott et al.'s study (2012), the target construction described an approach event ("To the Y goes the X"), denoting that X approaches Y in a certain manner: *To the gorilla jumps the camel* or, in German, "Zum Gorilla hüpft das Kamel. " Note that this construction is grammatical in German, though marked.[4] In English, albeit felicitous, this syntactic pattern is rare, potentially limiting transfer effects. Similar reasoning to limit transfer effects can be found in Viviani et al. (2026), Gómez (2002) and Wonnacott et al.'s (2012) studies. While Gómez (2002) and Wonnacott et al. (2012) used artificial language input, Viviani et al. (2026) used Japanese patterns that do not exist in English. In contrast, the current approach pattern is not entirely unknown to English speakers. German and English are both West-Germanic languages, so constructions with different global construction semantics are rare. While this makes avoiding transfer effects challenging, it resembles a natural FL learning phenomenon. In many FL classrooms, input patterns coexist in the L1 and the FL, especially if both languages share similar origins. In the current study, if there were potential transfer effects, they would be expected in both groups.

The target construction consisted of a prepositional phrase featuring the object (i.e, *zum Gorilla*) and of a verb phrase split into a movement verb (i.e, *hüpft*) and the subject (i.e, *das Kamel*). All children were exposed to exemplars in which the nouns in the object and subject positions varied, however, children in the no-variability condition encountered only one verb in the verb slot, while the variability condition featured four verbs. To ensure students understood the rest of the construction, all nouns in object and subject position were animals which were German–English cognates. The animal context was chosen to match Wonnacott et al.'s (2012) study and was a topic that would have been covered in German lessons anyway. Following Wonnacott et al. (2012)'s design, a set of six animals was selected and these were represented by six toy animals. The final selection included a *camel* ("Kamel"), *gorilla* ("Gorilla"), *bear* ("Bär"), *elephant* ("Elefant"), *zebra* ("Zebra"), and *giraffe* ("Giraffe").

The verbs selected for the verb slot also needed to meet several criteria. First, each verb had to denote an approach event and have the same number of syllables. Second, the verbs had to be new to students, without prior exposure in the curriculum. Third, while all verbs needed to represent approach events, their meanings had to be distinct enough to be clearly enacted with toy animals, avoiding ambiguities (e.g., the subtle difference between *bouncing* and *jumping* would be hard to act out precisely). Verbs related to animals' natural movements, such as *walking* or *galloping*, were excluded to avoid potential biases based on familiarity with animal behaviour. To ensure uniform difficulty, only verbs that were unnatural for all included animals were selected. After careful review and consultation with the school, the final set included: *jump* ("hüpfen"), *slide on your belly* ("rutschen"), *scoot* ("robben"), *tumble/do a somersault* ("purzeln"), *roll* ("rollen"), *sneak* ("schleichen"), and *slide on your back* ("schlittern"). Video demonstrations of each verb can be found in Appendix S2 in the Supporting Information online. While exact crosslinguistic semantic accuracy for verbs like *schlittern* and *rutschen* could be debated, the focus was on ensuring clear, distinct enactments. Since the children were likely unfamiliar with these verbs, and would not be explicitly taught their meanings, precision in enactment took precedence over linguistic nuance. Seven verbs were used following Wonnacott et al. (2012)'s design: four for training (though only one was used in the no-variability condition) and three reserved for testing (one for each of the three outcome measures).

Before the intervention, 16-sentence exposure sets were randomly generated for each condition, based on Wonnacott et al.'s (2012) study. Sixteen sentences per day (96 sentences total over six days) was deemed feasible by the host teacher. Using balanced randomization, 16 unique animal combinations were selected from the pool of six animals, resulting in each animal appearing five times, except giraffe and elephant, which appeared six times ($[4 \times 5 + 2 \times 6] \div 2 = 16$). Each combination featured two different animals, randomly assigned to either approach or be approached. For the variability group, four verbs were randomly selected from the verb pool and then randomly paired with the 16 animal combinations to create 16 sentences. For the no-variability group, one verb was randomly selected and combined with the 16 animal combinations, resulting in two distinct sets of 16 sentences, available in Appendix S3 in the Supporting Information online. In Wonnacott et al.'s, 2012, study, different children in the no-variability condition received different verbs as the single training verb, which was not feasible in our plenary teaching context. This will be addressed in the Discussion. Six different and randomized versions of each set were created to avoid ordering effects and used as input in each condition across the 6-day intervention.

Outcome Measurements

The outcome testing included three tasks—an act-out, production, and forced-choice task—following Wonnacott et al.'s design (2012). Each task involved both familiar and unfamiliar (novel verb) trials. In the act-out task, children listened to a sentence and enacted its meaning with toy animals. In the production task, they observed an enacted sentence and then produced the sentence. The forced-choice task presented two videos with reversed linking rules, and children selected the correct one based on a sentence prompt. Test sentences featured the same six animals from the exposure phase. Each task consisted of four trials and was completed twice, once with familiar verbs and once with unfamiliar verbs. Thus, each child completed 12 familiar and 12 unfamiliar trials across the three tasks. Accordingly, two sets of 12 animal combinations each were generated, ensuring balanced randomization of animal frequency and order. For each child in the variability group, one of the four familiar verbs from the exposure phase was randomly paired with the first 12 animal combinations to create familiar test sentences. Thus, during familiar trials, the children were tested on one familiar verb across all three tasks, like was done in Wonnacott et al.'s study (2012). In the no-variability condition, the single exposure verb was used in the familiar trials and also across all three tasks. For unfamiliar trials, three verbs that were unused during exposure in the variability condition

Table 2 Time frame intervention study

	Pre-Intervention	Intervention						Posttests		
Day	two weeks prior	1	2	3	4	5	6	7	8	9
Class 1	background tests	variability condition						testing		
Class 2	background tests	no-variability condition						testing		

were randomly paired with the remaining set of 12 animal combinations. The same three unfamiliar verbs were used for testing children in both the no-variability and variability conditions, even though there were technically six possible unfamiliar verbs for the no-variability condition. To avoid verb effects during testing, the identical, unfamiliar verbs were used for all children, regardless of input condition. Unlike the familiar trials, each of the three unfamiliar verbs was randomly assigned to a specific task for each child, ensuring one unfamiliar verb per task. Appendix S4 in the Supporting Information online provides sample test sentences for a child in each exposure condition. This process was repeated for each child in both input conditions. While the act-out and production tasks were enacted in person, the forced-choice task was tested using a computerized task built in the Gorilla software (Anwyl-Irvine et al., 2020). In each trial, children saw two videos on-screen. Both showed the same approach event with the same two animals, but only one video showed the correct linking rule; the other showed the reversed (incorrect) linking rule. The position of the correct video (top or bottom) was randomized per trial.

Procedure

The study followed the timeframe shown in Table 2.

Before the intervention, we coordinated with the host teacher to structure how the target sentences would be introduced in daily German lessons, balancing authentic classroom interaction with controlled input delivery. Pedagogic authenticity—the crux of the current experiment—was crucial to simulate typical teaching. On the other hand, the controllability of the input and the comparability of teaching settings across conditions was crucial from an experimental perspective. Thus, while we aimed to balance teaching authenticity with rigorous experimental control, we prioritized control in this study, given that it serves as a proof of concept to identify whether this type of input manipulation can be effective in FL classrooms. Due to frequent student absences (e.g., for music lessons), we decided to dedicate a 15-min segment at the start of each lesson to deliver the input uniformly across classes. This timing allowed all students to be present, addressing the interruptions noticed during preliminary classroom observations. During each 15-min segment, the first author—a trained FL teacher—conducted the teaching while the host teacher remained in the room. Children gathered in a circle, and the first author used a lapel microphone to record their teaching input, ensuring consistency by adhering to a printed script, with recordings verifying that each lesson followed the intended input delivery without omissions.[5]

At the start of the intervention, students in both classes were introduced to the six toy animals to ensure familiarity with their names. This included role-play activities where students "purchased" the animals and bargained over price. Throughout the intervention's six days, the first author, in their role as teacher, spent a few minutes daily exploring topics related to the animals, such as their origins (using an inflatable globe), weight (using a scale), prior appearances in books and movies (on National Book Day), and diet (using hay and grass). These introductions, which were conducted in English and lasted about five minutes, aimed to maintain engagement without using the target language patterns. We would then smoothly transition into the day's target input sentences (i.e., approach–events), with phrases like "Now the animals visit each other" or "Now they invite each other over for dinner" ensuring consistency across both classes. After each transition, the daily exposure phase began, which included 16 target sentences. For each sentence, the teacher first read it aloud while enacting it with the animals. The teacher then had the children repeat it together as a group while the teacher enacted the sentence again. This setup provided controlled, yet authentic, primary school classroom engagement in the form of call-and-response. Once all sentences were covered, students took a short break, and their regular German teacher resumed the lesson without further animal references to control for exposure frequency in each class. Note that all students were present for all six sessions with no absences.

Outcome Measurements Procedure

All background tests and outcome measures were piloted, with details in Appendix S5 in the Supporting Information online. Outcome testing began on Day 7, following six exposure days, and was conducted individually in the mornings to avoid conflicts with afternoon school activities. Testing took place in a prepared corridor space, with toy animals and a laptop for the forced-choice task. Teachers released students individually to minimize classroom disruption. Testing alternated between the variability and no-variability conditions. Although this rotation made logistics more challenging, it was implemented to ensure that students from one input condition were not tested one or two days ahead of those from the other condition. Testing took 15–20 min per student. Most students completed testing within the first two days, with only two students finishing on Day 3 due to absences. Each child's testing session began with casual conversation to help them feel comfortable. Then, the testing, which followed a consistent script, began. First, the act-out task, the production task, and the forced-choice task were conducted with one familiar verb each. These tasks were then repeated in the same order, using one unfamiliar verb each. This sequence was chosen to ensure children performed well on familiar tasks, which would start the testing positively and help maintain motivation. The act-out task was done first, as it was considered the most engaging, allowing children to interact with the animals. To minimize switching between animals and the laptop, the production task followed, with the forced-choice task last. The order of tasks was consistent for each child. Detailed procedures are available in Appendix S6 in the Supporting Information online.

Note that, with both the act-out and production tasks, it was necessary to have a way to teach the child the meanings of the unfamiliar verbs. We used the same method as Wonnacott et al. (2012) did.[6] Details are provided in Appendix S6 in the Supporting Information online.

Throughout the testing process, children were not given feedback on the accuracy of their responses. Regardless of their answers, they were consistently praised and encouraged to continue their good work.

As video recordings were not part of this study, all coding across all three outcome measures was completed by the main researcher. Based on experiences from piloting, they used a predefined error list with assigned numbers to quickly log responses during testing, allowing them to focus on the task and child interaction without writing detailed error descriptions. Incorrect grammatical gender was not considered an error as it rarely impacts meaning comprehension in German.

Statistical Approach

Separate analyses were conducted for each outcome measurement task. All analyses were conducted in the R Computing Environment (R Core Team, 2021). Data, materials, and script are provided here https://osf.io/9zfgu/overview. For each data set, we ran analyses with a binary dependent variable, indicating whether children made a correct/incorrect response. Overall, two separate sets of analyses were conducted over two different types of response coding: In the first coding type, responses were scored "correct" when the linking rule was applied correctly, regardless of verb accuracy. All other responses, mainly consisting of wrong target approach order, were coded as "incorrect." Unlike Wonnacott et al. (2012), who additionally considered responses where the approach event itself was correct, but the approach order was not, our first coding type was determined by the older age of our participants. Whereas Wonnacott et al. (2012), who studied significantly younger children (mean age 5;2), found that 55% of responses in their "one verb unfamiliar" trials fell into their other category, our pilot data did not suggest that 6-to-7-year-olds would fail to acquire the global approach semantics. Thus, our first coding type was concerned with whether participants had generalized the entire construction, involving both the construction's global approach feature (i.e., an approach event is happening) and the target approach order (i.e., animal B approaches animal A, and not vice versa).

Given that the pedagogical background of this work specifically targeted young FL students' productive skills, in the second coding type, responses were scored as being correct only if a participant produced an entirely correct approach event, including the target approach order and the verb semantics. This second coding type focused on the idea that simply getting the syntactic structure of a phrase right is typically not sufficient in an authentic FL communication situation, especially if the semantics of the featured verb are wrong. All remaining responses were scored as being incorrect.

Data were analyzed using a method which combined building (frequentist) mixed effect models and Bayesian inference using Bayes factors (Silvey et al., 2025; BF). For each data set, we built a generalized linear mixed model (GLMM) of the binomial family with a logit link function predicting response accuracy. All models had fixed effects of condition (variability vs. no-variability) and familiarity at test (familiar vs. unfamiliar) as well as their interaction. An additional independent variable for test day, reflecting whether the child took the test either on Day 1, Day 2 or Day 3 after exposure, was also included since this generally contributed to the model. However, test day was considered a control variable (acting like a covariate in ANCOVA); therefore, no interactions with this variable were included. It was included as a fixed effect to control for systematic differences across the three testing days shared by all participants, such as fatigue or day-specific effects (e.g., unrelated events during the school day that may have affected concentration). All predictors were centred such that the intercept represented the grand mean and the fixed effects can be interpreted as main effects. Random intercepts for participants, a random by-participant slope for familiarity at test, and correlations between the two were also included. In each task, additional models over subsetted data were run for each of the familiar and unfamiliar verb trials. These models were equivalent but possessed just one fixed effect (condition) and random intercepts for participants. Note that we assessed whether background test scores correlated with outcome measures to consider their inclusion as covariates. As detailed in Appendix S7 in the Supporting Information online, no clear associations were found, so background test scores were not included to avoid overcomplicating the models.

Statistics taken from the coefficients for the fixed effects were used to test a set of hypotheses which came from the assumption that variability promotes generalization while high token frequency promotes stronger "whole unit" learning with those items. These hypotheses, and how they relate to the effects in the different models, are available in Appendix S8 in the Supporting Information online. Each coefficient that was extracted provided an estimate of the difference in question (*b*) and its associated standard error (in log odds space). The associated *z*-scores and *p*-values were also automatically provided; however, in the current work, instead of basing our inferences on frequentist *p*-values, we instead computed and interpreted a BF for each of these effects. The BF quantifies the strength of evidence in favour of one hypothesis (H_1) over another (H_0), or vice versa. This method allowed researchers to precisely measure the degree of support for a given hypothesis (H_1) when compared to the null hypothesis (H_0). This stands in contrast to *p*-values, which lack the capability to assess the level of evidence in favour of the null hypothesis ($p > .05$ results do not signify increased confidence in the null hypothesis).

BFs test if the data is more likely under H_1 or H_0, and thus its computation requires a model of the plausibility of different effects under H_1 and a model of H_0. Following the work of Dienes (2014), H_1 was modelled as a half-normal distribution (one tailed, because predictions are directional) with a mean of 0 and the standard deviation (*SD*) set to a rough estimate of the predicted difference for the hypothesis in question. H_0 was modelled as a single point null. This computation requires three numbers: (1) an estimated mean difference for the effect in the data (*b*), (2) the associated standard error (*SE*), and (3) a rough estimate of the predicted mean difference (*s*) for that effect under H_1. For each BF calculation, *b* and *SE* were drawn from the relevant coefficients in the respective mixed effects models.[7] For the predicted mean differences, it is ideal for values from previous similar studies to be used. These were thus taken from Wonnacott et al.'s (2012) study wherever possible, that is, wherever an equivalent effect was indeed found in an equivalent, or similar, analyses for that task. In cases where previous values were not available, a plausible maximum was used to determine a plausible estimate of H_1, following the work of

Silvey et al. (2025). The details are presented in Appendix S8 in the Supporting Information online.

BFs are continuous, unlike *p*-values, enabling nuanced interpretation. They are often categorized into evidential categories (Dienes, 2014): BF $>$ 10 strongly supports H_1, BF $>$ 3 indicates moderate/substantial support (often aligning with $p < .05$), BF $<$ 0.33 moderately/substantially supports H_0, and BF $<$ 0.10 strongly supports H_0. Ambiguous evidence lies between these ranges. Robustness regions (RR) were calculated to account for our subjective H_1 estimates. These are a type of sensitivity analysis, showing the range of estimates of H_1 for which our data would support the same conclusion (i.e., evidence for H_1, evidence for H_0, or inconclusive, based on the cut-offs of BF $>$ 3 or BF $<$ 0.33). The larger the robustness region, the wider the range of values of H_1 estimates that would support the same conclusion. RRs are noted as [x1, x2], representing the smallest (x1) and largest (x2) compatible standard deviation (Dienes, 2021).

In addition to the main analyses, an error analysis will be presented to provide further insight into transfer effects and verb accuracy.

Results

Statistical Analyses

Corresponding to our two coding types, two overviews of participants' mean response scores across all three task types are provided in Figures 1 and 2.

For both types of response coding, the plots indicate that the no-variability group outperformed the variability group during testing with familiar items (although with only small mean differences), whereas the reverse was true for unfamiliar trials with considerably larger mean differences between groups. This latter effect was confirmed statistically (see Results below). Note that we acknowledge the variability shown in the plot (see Discussion), but we emphasize that our key statistics concern the group differences.

Act Out

Statistics for our analyses on linking rule correct responses are in Appendix S9 in the Supporting Information online. There was substantial evidence for an interaction between input group and familiarity with BF = 7.6 (predicted effect = 3.93, RR = [2; 110], $b = 11.59$, $SE = 4.28$, $z = 2.71$, $p = .0068$) indicating that the data was more than 7 times more likely to fall under H_1 than under H_0. This interaction was broken down statistically by testing evidence for (a) the hypothesis that there would be better performance in the no-variability than in the variability group for familiar verbs and for (b) the hypothesis that

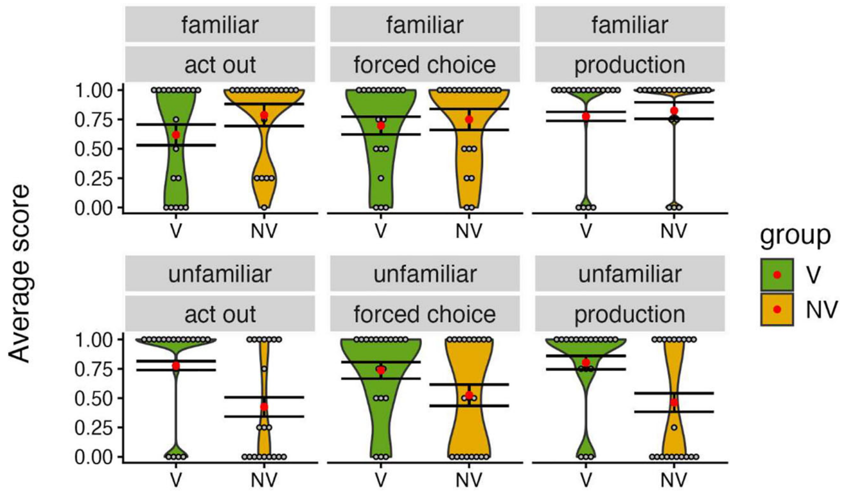

Figure 1 Violin plots displaying participant mean proportion correct in each of the three task types during outcome measurements where correct scores are responses where the global construction semantics and linking rule are applied correctly, regardless of verb accuracy. V = variability. NV = no-variability. Error bars present 95% confidence intervals.

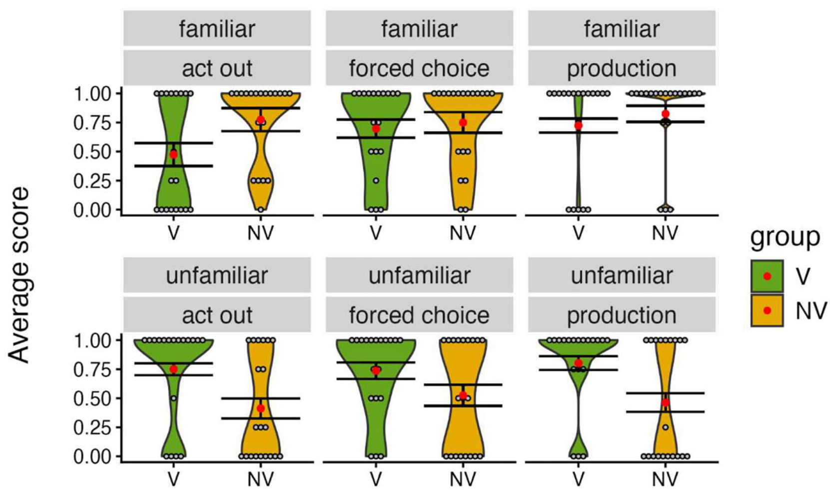

Figure 2 Violin plots displaying participant mean proportion correct in each of the three task types during outcome measurements where correct scores are responses where the global construction semantics, the linking rule, and the verb semantics are applied correctly. V = variability. NV = no-variability. Error bars present 95% confidence intervals.

there would be better performance in the variability than in the no-variability group for unfamiliar verbs. Evidence for the former, (a), was ambiguous with BF = 0.72 (predicted effect = 6.08, RR = [0; 16.8], $b = -1.9$, $z = -0.72$, $p = .47$). Evidence for the latter, (b), was substantial with BF = 3.59 (predicted effect = 1.52, RR = [1.4; 5132], $b = 18.43$, $z = 3.96$, $p < .001$). Thus, the data is more than 3 times more likely under H_1.

Analyses on entirely correct responses (see Appendix S10 in the Supporting Information online) confirmed that there was strong evidence for an interaction between input group and familiarity with BF = 15 (predicted effect = 3.93, RR = [1.6; 330], $b = 12.36$, $SE = 3.98$, $z = 3.11$, $p = .0019$) indicating that the data was 15 times more likely under H_1 than under H_0. When this was broken down by verb familiarity, evidence for better performance in the no-variability than in the variability group for familiar verbs was substantial with BF = 3.66 (predicted effect = 2.82, RR = [0; 2.1], $b = 5.91$, $z = -2.04$, $p = .04$) while evidence for the hypothesis that there would be better performance in the variability group than in the no-variability group for unfamiliar verbs was ambiguous with BF = 1.49 (predicted effect = 1.52, RR = [0; 6.3], $b = 9.50$, $z = 1.79$, $p = .07$).

Production

Statistics for analyses on linking rule correct responses are in Appendix S9 in the Supporting Information online. There was substantial evidence for an interaction in the predicted direction between input group and familiarity with BF = 368 (predicted effect = 3.93, RR = [0.9; 6641], $b = 20.02$, $z = 4.83$, $p < .001$) indicating that the data was more than 300 times more likely under H_1 than under H_0. This interaction was broken down statistically by verb familiarity. There was evidence against the hypothesis that there would be better performance in the no-variability group than in the variability group for familiar verbs, that is, substantial evidence for H_0 with BF = 0.28 (predicted effect = 10.71, RR = [0; 9.9], $b = -0.22$, $z = -0.08$, $p = .94$). There was substantial evidence for higher performance in the variability than no-variability condition for unfamiliar verbs with BF = 4.53 (predicted effect = 1.52, RR = [1.2; 6627], $b = 20.95$, $z = 4.47$, $p < .001$). Thus, the data were more than 4 times more likely to fall under H_1.

Analyses on entirely correct responses (see Appendix S10 in the Supporting Information online) confirmed that there was strong evidence for an interaction between input group and familiarity with BF = 1557 (predicted effect = 3.93, RR = [0.83;7549], $b = 21.75$, $z = 5.37$, $p < .001$) indicating that the data were more than 1,500 times more likely to fall under H_1 than under H_0. Breaking this down by verb familiarity, there was evidence against the hypothesis that there would be better performance in the no-variability than variability group for familiar verbs, that is, substantial evidence for H_0 with BF = 0.29 (predicted effect = 10.69, RR = [0; 10.65], $b = 0.51$, $z = -0.18$, $p = .86$). There was substantial evidence for the hypothesis that performance was higher in the variability than no-variability group for unfamiliar verbs with BF = 4.53 (predicted effect = 1.52, RR = [1.2; 6626], $b = 20.95$, $z = 4.47$, $p < .001$). Thus, the data were more than 4 times more likely under H_1.

Forced Choice

Figures 1 and 2 generally indicate that the variability group outperformed the no-variability group during testing with unfamiliar items and that the reverse was true for familiar trials. However, the statistics relevant for testing the hypotheses that children in the variability condition would show better performance on unfamiliar items in the forced-choice task than would children in the no-variability condition (see Appendix S11 in the Supporting Information online) provided only ambiguous evidence for an interaction between input group and familiarity with BF = 1.3 (predicted effect = 5.1, RR = [0; 25.2], $b = 3.81$, $z = 1.83$, $p = .07$). Broken down by verb familiarity, the evidence for the hypotheses that that there would be (a) better performance in the no-variability than variability group for familiar verbs and (b) better performance in the variability than no-variability group for unfamiliar verbs was also ambiguous in each case: (a) BF = 0.42 (predicted effect = 3.2, RR = [0; 4.6], $b = 0.11$, $z = 0.07$, $p = .95$); (b) BF = 2.2 (predicted effect = 1.9, RR = [0; 90], $b = 5.34$, $z = 1.76$, $p = .08$).

Summary

Across two of three tasks, the descriptive data showed a trend of stronger performance in the no-variability group for familiar items and in the variability group for unfamiliar items. Statistical analyses of the act-out and production tasks using both linking rule correct and entirely correct coding strongly supported an interaction between input condition and verb familiarity. Focusing on the benefits for the variability group for generalization, when responses were coded as correct, focused solely on how accurately thematic roles were mapped to syntactic positions (regardless of verb accuracy), both tasks showed substantial evidence for stronger performance in the variability group than the no-variability group in the unfamiliar verbs task. This suggests that input variability supported generalization of linguistic patterns to novel verbs in this experimental context. For the analyses where the verb also had to be accurate,

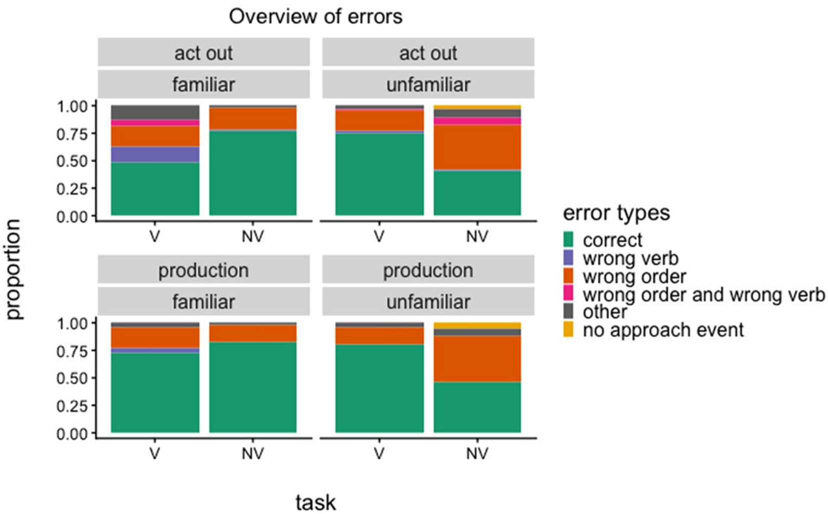

Figure 3 Distribution of different response types during outcome measurements displayed by task type, condition, and familiarity. V = variability. NV = no-variability.

we again found substantial evidence for the variability benefit in the production task, but the evidence in the act-out task was ambiguous.

Results for the forced-choice task also suggested a no-variability benefit for familiar items and a variability benefit for novel items, though evidence for the interaction remained ambiguous. For unfamiliar trials, the effect of input variability on linking rule accuracy was in the positive direction but remained ambiguous.

Error Analysis for Act-Out and Production Tasks

The types of errors that participants were making were further investigated. The categories were:

- fully correct
- correct order, wrong verb (correct linking rule but wrong verb semantics)
- wrong order and wrong verb (linking rule and verb semantics wrong)
- incorrect approach event (approach event but wrong animal(s) and/or wrong order and/or wrong verb)
- no approach event (child produced/acted out an unidentifiable action)

Figure 3 provides an overview of the distribution of the different errors.

The analysis revealed that most errors stemmed from using the wrong linking rule as most errors made by participants in both groups and across both types of tasks and familiarities were wrong order errors (orange). This error type, likely a transfer effect, is considered in the Discussion. Other error types accounted for only a smaller amount of the general distribution of error types. An important observation is that children generally understood the construction's global approach semantics. This contrasts with results from work with younger children (Wonnacott et al., 2012).

The largest proportion of wrong verb responses (purple) came from participants in the variability group tested on familiar verbs in the act-out task, potentially reflecting the fact that this group encountered each individual verb less frequently than the no-variability group. The phenomenon of getting the linking rule right but using the wrong verb, and vice versa (i.e., getting the verb right but applying the wrong linking rule), is discussed below.

Discussion

In the current work, 6-to-7-year-olds learned a new VAC in German. Building on an earlier experimental study (Wonnacott et al., 2012), we compared the learning of two groups of children trained with either variability (i.e., construction exemplified with multiple verbs) or no-variability (i.e., construction exemplified through single verb) but embedded this into a classroom context. The findings largely supported the previous work, with children learning in all conditions but, at least in two of the tasks, there was more generalization in the variability group. For the remainder of this discussion, we draw out some key aspects of the findings and discuss relevant differences with the previous work.

Overall Learning and the Input Paradigm

Before discussing the key variability manipulation, we want to highlight that students across conditions learned something about the target pattern in the input without receiving any explicit rule instruction. Importantly, the current experiment used natural language input, meaning that children had to navigate input that featured authentic—and potentially challenging—linguistic characteristics, such as German phonological characteristics (while, for example, Wonnacott et al., 2012, used linguistic stimuli with English L1 morphological and phonological characteristics). Moreover, this learning occurred despite the challenges posed by the classroom context. This corroborates what curricula have begun to highlight; namely, that young, beginner FL learners can pick up on "large" linguistic units in the input and, at the very least, already recognize or imitate them after a relatively small amount of exposure.

The Verb Type Frequency Effect in the Noisy Classroom

The results of the analyses for the act-out and production tasks with novel verbs provide good evidence that children in the variability condition showed a stronger ability to generalize the abstract linking rule of the construction to novel verbs than those in the no-variability condition. It is notable that, despite the noisier classroom context, these results offer additional clarity when considered with those of Wonnacott et al. (2012), who found no evidence of a variability advantage in the equivalent analyses for production tasks where the linking rule had to be applied correctly (though, recall they did find a variability advantage in learning the global meaning of the construction). We believe this difference is most likely due to the differing ages and cognitive developmental stages of the participants (4-to-5-year-olds in Wonnacott et al., 2012; 6-to-7-year-olds here). This conclusion is further supported by the fact that Wonnacott et al. (2012) reported 55% (act-out) and 49% (production) failure rates in learning approach semantics, whereas such failures were rare (< 1%) in our study. Note, however, that finding the reported effect in the current classroom context might have benefited from the absence of noise stemming from peers' inconsistent productions (cf. Hopp, 2016) given the restrictive call-and-response input format without peer-to-peer interactions.

Turning to the analyses focused on entirely correct responses (i.e., including verb accuracy), we again found substantial evidence for a variability benefit for unfamiliar verbs in the production task, but the results for the act-out task were ambiguous. Inspecting the data, it is notable that only three act-out responses by students in the variability group were incorrect, highlighting the small sample size on which these conclusions are based. While we can thus not draw any conclusion from this part of the analysis, it is worth noting that, while variability will support learning of the pattern as it promotes schematization and productive pattern extension, it will not necessarily support the ability to copy the action correctly, which was also tapped into via this stricter coding. Thus, we believe that the data provide convincing evidence across two tasks that the ability to infer the underlying linking rule and generalize it to novel contexts was stronger in the variability group than the no-variability group. This finding is broadly consistent with other findings in the literature (Raviv et al., 2022; Viviani et al., 2026; Wonnacott et al., 2012), providing preliminary evidence that the verb-type frequency effect can also hold in the noisy environment of a primary FL classroom. This is good news because it is precisely young FL learners' apparent lack of structural representation of language that has been criticized in the literature, resulting in calls to find ways to equip students with a productive repertoire of linguistic patterns and help them overcome the imitative plateau.

Variability and Verb Accuracy

Learning about underlying structural patterns came at the cost of verb accuracy, as demonstrated by the struggle of students in the variability group with verb accuracy in familiar trials compared to the students in the no-variability group. This is clearest in Figure 3, where we can see that, in familiar trials, students from the no-variability group almost never got the verb wrong even though they frequently got the linking rule wrong. This is unsurprising given that each individual verb is encountered considerably less frequently in the variable input than the single verb in the invariable input (i.e., decreased token frequency). It nevertheless demonstrates that there is a trade-off between variability and token frequency, which is an important caveat to the positive benefits of variable input (cf. Endress & Hauser, 2011). That said, although we expected this conclusion to be further supported by stronger performance in the construction learning measures for unfamiliar verbs, the evidence here was rather mixed.

From a pedagogical perspective, as the students' German teacher highlighted during the study, there does not seem to be a point for the variability group in gaining structural knowledge without getting the verb correct.[8] In theoretical terms, constructions may exist independently of individual verbs, but, in pedagogical contexts, individual students gain limited benefit from acquiring a construction abstractly (which is already a considerable achievement given the limited input) without sufficient verb knowledge to actually use or "activate" this knowledge. Given this, in the context of the current experiment, how could the lack of verb accuracy with familiar items in the variability group have been avoided? It might seem most straightforward to hand students who had figured out the pattern a list of appropriate verbs for each slot which was another (understandable) perspective raised by teachers during the study. However, in keeping with the usage-based, input-driven context of this work, an alternative would be to increase exposure with the goal of increasing implicit learning of the target verbs. Thus, we argue that the potentially detrimental impact of variability on verb learning in the variability condition could have been mitigated against by simply increasing overall exposure. In the variability group, the four individual verbs were more difficult to keep track of as their individual frequency was low, and students needed to track the four different movements (i.e., types of approach) associated with the four different verbs. By increasing the overall amount of exposure, however, the frequency of each individual verb increases as well, and, over time, students should be able to learn their meaning through repeated encounters with the verb and its associated meaning. Crucially, the equivalent cannot be said of learning of the pattern in the no-variability condition. That is, even if the invariable input was frequently repeated, the underlying pattern of the VAC might be less likely to be learned in the no-variability group because the reliability of the cues regarding the characteristics of the underlying pattern would not increase and, therefore, remain comparatively weak (i.e., compared to the variability condition). However, this remains an empirical question. In the no-variability condition, the input provides only weak—albeit repetitive—evidence for the link between the subject–verb–PP pattern and patient–goal semantics and no evidence that it is applicable independently from the intervening verb. With only one intervening verb (i.e., *hüpft*), the construction's structure is never encountered without this specific verb. In an input-driven, constructionist view of language learning, this restrictiveness would suggest to learners that the construction can only be used with a limited number of verbs, or potentially only with this individual verb (Bybee, 1995). Crucially, an increase in exposure would not change anything about this issue, but a change in distribution might, see Skewed Input below. In contrast, verb learning in the variability group could improve over time with more input (i.e., as each verb increases in frequency in their sample).

Of course, there are constraints as to how many times input can reasonably be repeated in instructed settings. Simply increasing exposure in a scenario with increased input variability is not viable in most teaching contexts. In pedagogical contexts, this is most likely where explicit input and teacher feedback would naturally come into play.

In addition, learning verb semantics could be supported by using prototypical verbs with broad and flexible meanings, which would, in turn, allow for high-frequency use. While we did not find evidence for a prototypicality effect (i.e., no evidence of verb-specific performance differences), in future studies, it might be useful to include a prototypical verb in the target construction as a baseline to ensure that all children learn one broadly applicable verb. Introducing such a prototypical verb (e.g., *give* in a double-object construction) would necessarily suggest a skewed input distribution, which we consider below.

Skewed Input

An alternative input manipulation that could address the limitations of our variability condition (i.e., low verb accuracy) and no-variability condition (i.e., limited potential for forming analogies across exemplars, developing categories, and gradually inducing abstract patterns) is skewed input (e.g., Casenhiser & Goldberg, 2005; Pulido, 2024), which typically corresponds to a Zipfian distribution. This would avoid the drawbacks of using either a single verb type or many equally distributed verbs. Instead, a Zipfian distribution would allow one to keep the number of verb types constant across conditions (e.g., four) while varying input frequency.

As demonstrated in previous classroom research with adult learners (e.g., Madlener-Charpentier, 2015), skewed input helps learners by combining frequent, central exemplars with varied contexts, making these items highly salient and entrenched. These central exemplars serve as anchors for forming abstract patterns and categories, while additional variation in low-frequency items promotes the development of generalized schemas and category representations (cf. Barđdal, 2008; Madlener-Charpentier, 2015). On the other hand, note that Viviani et al. (2026) found evidence for no benefit of skewed input with primary school L2 learners, although this might be because they did not combine skewness with a prototypicality manipulation.

If the current study had included a skewed input condition, verb accuracy may have been expected to increase (compared to our variability group), at least for the single high-frequency (and ideally prototypical) verb, through a stronger entrenchment process during frequent exposure. On the other hand, because the number of intervening verbs would be greater than one, as opposed to our current no-variability condition, the cue strength for learners indicating that the verb can be dissociated from the construction's structure would increase. Learners could then form analogies across exemplars, recognizing that, although the verb changes, the underlying form–meaning link (e.g., the first named animal approaches the second in a specific way) remains constant. However, this remains speculative and warrants follow-up studies with young learners.

Forced-Choice Task

Although the means in the forced-choice task trended in the predicted direction of stronger performance in the variability group for unfamiliar items and stronger performance in the no-variability group for familiar items, we did not find substantial evidence for either effect. Since we also did not find evidence for the null hypothesis (i.e., BF analyses found that evidence was in the ambiguous category), any discussion of this data must be tentative. Nevertheless, we briefly consider why the verb type frequency effect might not be as strong in this task as in the other two tasks. One potential reason is the fact that this task was always the last task during testing. It may be that students' attention had decreased following the act-out and the production tasks, which were challenging and required high levels of concentration and effort from the children. Another possibility is that the task may have been too challenging. Participants were asked to view and choose between two videos showing the same approach action in two reverse orders while hearing the input sentences (half of them with unfamiliar verbs). This demanded high levels of attention. In addition, as a form of encouragement, students had been praised for all their responses in the previous tasks, regardless of accuracy. It remains unclear if students used this praise to adjust what they had inferred from the input during the previous exposure days or if this potential readjustment impacted their performance in the last task.

Interestingly, Wonnacott et al. (2012) also found no evidence of verb type frequency effects in their forced-choice task. They accounted for this result by suggesting an agent-first bias (Wonnacott et al., 2012) among English L1 speakers which corresponded to the typical pattern in their L1. This bias might have competed with the learned linking rules, resulting in ambiguous data. As further evidence of this they ran a second study in which children were exposed to the construction without its semantics, i.e., with no opportunity to learn the linking rule, and then gave them the same tests. Children showed a strong bias to pick the picture in which the first mentioned noun approached the second mentioned noun. A similar bias could also have impacted the results in the current experiment. While this complicates the interpretation of the forced-choice data, it is crucial to consider biases like this when analyzing data from FL teaching interventions that involve natural language input. In authentic teaching situations, FL learning is likely to be influenced by crosslinguistic transfer effects (Ballinger et al., 2020).

Crosslinguistic Transfer Effects

The wrong order errors, where children interpreted the first noun as the agent of the action (Figure 3), were most likely caused by transfer from English, which is to be expected given the close typological proximity of German and English, especially regarding the construction at hand. In instructed contexts using natural languages, typological distance is an influential variable that cannot be easily controlled for. In such cases, transfer effects, likely caused by entrenched similar patterns in the source language, should not be underestimated (cf. Jaekel et al., 2024). The question is how transfer effects might facilitate generalization in contexts with typologically similar languages, or how they might impede (or even block) generalization in contexts with typologically distant languages—especially when input manipulations are purely implicit, and the structural rules of the target language are not made explicit. This is an interesting issue for future research and an unavoidable issue for any FL classroom with a linguistically heterogenous student body. Currently, to our knowledge, the role of typological distance in FL generalization specifically from manipulated input in primary classrooms has not yet been examined in the experimental literature.

Explicit Input

In line with previous, more controlled research, the current study examined children's ability to infer structural information without explicit instruction. While primary FL classrooms often avoid extensive rule teaching to maintain motivation, some low-level explicit input (e.g., basic rule instruction, feedback, L1 translation) is typically provided (Legutke et al., 2012). A plethora of research has shown that such explicit instruction can facilitate FL learning, though this is mediated by variables such as practice type, practice duration, and delivery method (Akakura, 2012; Li & Sun, 2024). Explicit input in the form of corrective feedback has also been shown to be impactful in children's FL development (Lyster et al., 2013). For intervention studies like ours, it would be valuable to explore the impact of additional explicit input on learning outcomes across input conditions. For example, explicitly drawing students' attention to the target construction at the beginning of the exposure phase with prompts like, "Notice what repeats and changes in these sentences" might be beneficial for both groups, especially the variability group. Alternatively, if researchers provided simple rule instruction, such as identifying the first animal as the one being approached, it would be interesting to see whether, and to what degree, subsequent variability manipulations impact on the long-term fostering and development of this explicitly introduced structural knowledge foundation, a point already raised elsewhere (Madlener-Charpentier, 2015). For pedagogical purposes, it would also be useful to assess how input variability and continuous corrective feedback influences students' ability to apply structural knowledge independently and supports the development of their mental representations of form–function pairings.

Individual Differences

Figures 1 and 2 show considerable individual differences in performance. Particularly notable are (a) some children in the variability group failing to extend the form–meaning pairing (i.e., the target construction) to unknown verbs at test, and (b) some children in the no-variability group who seemed to generalize well, contrary to expectations. Currently, we lack additional data to explain these exceptions. Our background test battery (e.g., WASI, German/English language tests) did not reveal any relationship with performance on the outcome test. This might be the case because our included background measures were primarily access-related, given that they tapped into learners' ability to perceptually encode the input, rather than into abilities that directly impacted the underlying learning mechanisms once the input had been processed. One such access-related cognitive ability that may be relevant in the context of input frequency effects is auditory pattern discrimination, which has been found to interact with type frequency in adults (McDonough & De Vleeschauwer, 2012). Although not a mechanistic variable affecting learning processes per se, it is plausible that, in young learners too, auditory pattern discrimination ability interacts with type frequency by influencing how much of the more or less varied input is accessible for learning to happen in the first place.

In contrast, a variable that may impact learning mechanisms themselves is working memory. In usage-based approaches, *analogical comparison* (i.e., the ability to align and abstract relational structure across exemplars) has been shown to be influenced by working memory. In particular, children with lower working memory capacity benefit more from increased surface similarity across exemplars (Simms et al., 2018). This suggests that working memory may moderate the effects of different input conditions. For example, input with greater surface similarity (e.g., skewed input) may be particularly beneficial for learners with lower working memory capacity. In the present data, the successful learners in the variability condition may have benefited from relatively higher working memory capacity, although this was not measured and thus remains speculative.

Considerable individual differences (where even top performers may score low) are common in learning studies, though violin plots in our study make this variability more visible than traditional box plots. Nevertheless, from a pedagogical perspective, the fact that we did not observe variability benefits across the board at the group level is significant, given the heterogeneity of primary classrooms in FL proficiency and other individual differences. In heterogenous contexts, no-variability input might sometimes be a justified pedagogical choice. For example, Kostka (2020) investigated formulaic sequences in a classroom setting (though without a control group) and found that weaker FL learners relied, for longer periods of time, on unanalyzed, prefabricated input. This approach slightly expanded their communicative scope but in an imitative, rather than situation-independent, manner. In contrast, stronger students extracted structural information from the input, enabling them to increase their situation-independent communicative agency. As curricula rightly emphasize, input patterns can support weaker students by providing basic communicative tools while also serving stronger students as a resource for structural learning. However, the notion that no-variability input benefits weaker learners is not fully supported by our data. Despite our participants' marginal German proficiency, we observed limited no-variability advantages (mostly in verb accuracy).

Taken together, whether the learning goal is imitation or situation-independent production, whether individual differences are conceptualised as access-related or mechanistic, and whether studies focus on type frequency alone or also on distributional effects (e.g., skewing), future research in this area should consider aptitude–treatment interactions, as recently investigated by Pulido (2024) for adults. The interaction between input variables and learners' individual characteristics is likely to further inform our understanding of the mechanisms underlying learning from frequency-manipulated input and may offer promising directions for instructional practice.

Limitations

As part of the first author's doctoral project, this study faced limitations in sample size and characteristics due to organizational and financial constraints. For instance, the quasi-experimental design (relying on only two intact classes) raises the possibility that some observed group differences may be attributable to unmeasured variables (e.g., motivation), rather than the experimental manipulation (i.e., variability) itself. In the face of such constraints, note that our use of BFs has the advantage that it allows us to assess whether "null" results truly indicate evidence for the null, as opposed to evidence which is ambiguous due to sample size limitations. In general, BFs are expected to become more extreme as the sample size increases, and we accumulate more evidence about whether H_1 or H_0 is more likely. With a large enough sample, we would always cross our evidence thresholds in one direction or the other (though that sample may be infeasible). The current experiment explicitly serves as a proof of concept, providing a foundation for future, higher-powered studies. The parameter estimates obtained from the current dataset can also be used in simulations to determine how large future samples need to be.

Another limitation in the current study concerns the necessary trade-off between ecological validity and experimental control, particularly in the context of a study targeting usage-based language learning. Our goal of an authentic classroom setting conflicts with our concern for controllability. Usage-based approaches emphasize meaningful, communicative interaction as the foundation for experience-driven, input-driven learning. Yet, our call-and-response format is inherently more controlled than naturally occurring classroom discourse, even if, in many primary school contexts by virtue of the children‘s rudimental L2 proficiency, L2 input may often take a call-and-response form. It is therefore valid to ask how genuinely “meaningful” the students’ interactions in our study were. In addition, it is valid to question the generalizability of the findings to other interactive contexts in the classroom. Although we endeavoured to situate the target language in engaging scenarios (e.g., animals hosting each other for dinner or birthdays), participants were neither active agents in those interactions nor were they unaware of the study’s contrived nature. Indeed, in line with ethical regulations, each child provided informed consent, making them necessarily conscious of the unusual circumstances—an issue that likely applies across all experimental classroom-based research examining forms of implicit, incidental learning. Consequently, the very design and contextual constraints of our study unavoidably limit its ecological validity.

We wish to emphasize, however, that this investigation was conceived as a proof of concept. In line with numerous calls in the literature (e.g., Goldberg & Casenhiser, 2008; Madlener-Charpentier & Pagonis, 2022; Robinson & Ellis, 2008), we conducted a first test of whether we can manipulate primary school L2 input so as to promote generalization, even amid the inevitable everyday disruptions and distractions of a Year 2 classroom. We believe this explicit incorporation of the classroom’s literally “noisy” features represents a novel and important contribution to the field.

That said, we fully acknowledge that our study falls somewhat short of the level of authentic, learner-driven interaction advocated by usage-based theory. To address this, we are planning follow-up studies that will foreground task-based, game-like, intentional activities in which students (a) engage in genuinely meaningful interactions and (b) remain largely unaware of the study’s aims because the objectives are integrated into the gameplay. This would enhance ecological validity and thus provide a stronger test of the extent to which insights from cognitive science can be extended to authentic primary L2 classrooms.

Conclusion

Given the results of a systematic review of the extant research (Schulz et al., 2023), we believe that our intervention constitutes the first investigation into the impact of input variability on young FL learners’ learning and generalization of underlying form–function mappings of a VAC in an authentic FL classroom. Acquiring such generalizations is considered crucial for young FL learners' development of context-independent, productive communicative agency. However, FL learning outcomes in input-limited instructed settings are often suboptimal, with students frequently leaving primary school possessing relatively limited and largely only imitative language abilities.

The results of the current teaching intervention tentatively confirm the positive impact of input variability on students' ability to infer and generalize underlying structural patterns—even in a noisy classroom context. Specifically, the data focused on how accurately thematic roles were mapped to syntactic positions (regardless of verb accuracy) from unfamiliar trials in the act-out and production tasks. It yielded substantial BF evidence in favor of the hypothesis of improved performance by the variability group in contexts with unfamiliar verbs. This verb type frequency effect corresponds to previous, more controlled research in psycholinguistics but also across other cognitive domains. However, the current data are also in line with previous research in the way they indicate that variable input is no panacea and may, in certain situations and for certain learners, even be at best not beneficial to learning. In fact, the data suggest that no-variability input facilitates imitation and may therefore (continue to) play an important role in early FL pedagogy that should not be overlooked.

More generally, the current work illustrates that, even in a classroom, young FL learners can pick up on complex form–function correspondences from extremely limited input, provided it is purposefully structured. Tentatively, the data provide researchers and educators alike with the important confidence that, when input parameters related to variability are set correctly, young FL students' communicative agency may be improved, even in severely input limited contexts. Based on these promising proof-of-concept results, further high-powered interventions using a randomized controlled trial design are needed, focusing on the effects of different types of explicit teaching, varying type–token input ratios, and the mediating effects of individual differences.

Final revised version accepted 8 May 2026

Open Research Badges

This article has earned an Open Data and Open Materials badge for making publicly available the components of the research methods needed to reproduce the reported procedure. All data and materials that the authors have used and have the right to share are available at https://osf.io/5ayvr/. All proprietary data and materials have been precisely identified in the manuscript.

Notes

1 Note that these learners were likely more cognitively advanced than the 6- to 7-year-olds in the present study.
2 University of Oxford, Ethics Approval Reference: CIA-22TT-135
3 Expected WASI raw scores are lower than those observed here. This may be due to the unsupervised group setting reducing pressure, or because children revisited test items (despite instructions not to).
4 The construction is marked in German in that it represents an object–subject word order, whereas the preferred word order in German main clauses is typically subject–object (Bader & Häussler, 2010). In English, the preferred word order in main clauses is also subject–object (Crystal, 2018). This object argument topicalization is most likely in German when the object bears a relation to the preceding discourse, or in relative clauses with a personal pronoun as the subject (Bader & Koukoulioti, 2018).
5 Recordings cannot be shared due to consent requirements.
6 This was not required for the forced-choice task since it was not necessary to know the meaning of the verb to determine which scene depicted the correct linking rules.
7 This method assumes normality which is met since the calculations were in log-odds space (Brekelmans et al., 2022; Silvey et al., 2025).
8 Vice versa, there also seems to be no point in getting the verb correct but lacking structural knowledge.

References

Akakura, M. (2012). Evaluating the effectiveness of explicit instruction on implicit and explicit L2 knowledge. *Language Teaching Research*, *16*(1), 9–37. https://doi.org/10.1177/1362168811423339

Anthony, L., & Nation, I.S.P. (2021). PVST (Version 1.2.3) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Bader, M., & Häussler, J. (2010). Word order in German: A corpus study. *Lingua*, *120*(3), 717–762. https://doi.org/10.1016/j.lingua.2009.05.007

Bader, M., & Koukoulioti, V. (2018). When object-subject order is preferred to subject-object order: The case of German main and relative clauses. In, *Grammar and Corpora 2016* (pp. 53–72). Heidelberg University Publishing. https://doi.org/10.17885/HEIUP.361.C4692

Ballinger, S., Man Chu Lau, S., & Quevillon Lacasse, C. (2020). Cross-linguistic pedagogy: Harnessing transfer in the classroom. *The Canadian Modern Language Review*, *76*(4), 265–277. https://doi.org/10.3138/cmlr-76.4.001-en

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modelling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*(41), 17284–17289. https://doi.org/10.1073/pnas.0905638106

Barðdal, J. (2008). *Productivity: Evidence from case and argument structure in Icelandic* (Vol. *8*). John Benjamins Publishing Company. https://doi.org/10.1075/cal.8

Brekelmans, G., Lavan, N., Saito, H., Clayards, M., & Wonnacott, E. (2022). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *Journal of Memory and Language*, *126*, 104352. https://doi.org/10.1016/j.jml.2022.104352

Brown, H., Smith, K., Samara, A., & Wonnacott, E. (2021). Semantic cues in language learning: An artificial language study with adult and child learners. *Language, Cognition and Neuroscience*, *37*(4), 509–531. https://doi.org/10.1080/23273798.2021.1995612

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*, 425–455.

Bybee, J. (2008). Usage-based grammar and second language acquisition. In P. J. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 216–236). Routledge.

Casenhiser, D., & Goldberg, A. E. (2005). Fast mapping between a phrasal form and meaning. *Developmental Science*, *8*(6), 500–508. https://doi.org/10.1111/j.1467-7687.2005.00441.x

Crystal, D. (2018). *The Cambridge encyclopedia of the English language*. Cambridge: CUP.

Department for Education. (2013). *National curriculum in England: Languages programmes of study*. https://assets.publishing.service.gov.uk/media/5a7b9246e5274a7318b8f889/PRIMARY_national_curriculum_-_Languages.pdf

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2021). How to use and report Bayesian hypothesis tests. *Psychology of Consciousness: Theory, Research, and Practice*, *8*(1), 9–26. https://doi.org/10.1037/cns0000258

Doyé, P., & Lüttge, D. (1977). *Untersuchungen zum Englischunterricht in der Grundschule: Bericht über das Braunschweiger Forschungsprojekt „Frühbeginn des Englischunterrichts''*. Westermann.

Ellis, N. C. (2009). Optimizing the input: Frequency and sampling in usage-based and form-focused learning. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 139–158). Wiley. https://doi.org/10.1002/9781444315783.ch9

Ellis, N. C. (2017). Chunking in language usage, learning and change: I don't know. In M. Hundt, S. Mollin, & S. E. Pfenninger (Eds.), *The changing English language* (pp. 113–147). CUP. https://doi.org/10.1017/9781316091746.006

Ellis, N. C., & Ferreira-Junior, F. (2009). Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal*, *93*(3), 370–385. https://doi.org/10.1111/j.1540-4781.2009.00896.x

Ellis, N. C., & Wulff, S. (2020). Usage-based approaches to L2 acquisition. In B. VanPatten, G. D. Keating, & S. Wulff (Eds.), *Theories in second language acquisition: An introduction* (3rd ed., pp. 63–82). Routledge. https://doi.org/10.4324/9780429503986

Endress, A. D., & Hauser, M. D. (2011). The influence of type and token frequency on the acquisition of affixation patterns: Implications for language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 77–95. https://doi.org/10.1037/a0020210

Enever, J. (Ed.). (2011). *Early language learning in Europe*. British Council.

Engel, G., Groot-Wilken, B., & Thürmann, E. (Eds.). (2009). *Englisch in der Primarstufe - Chancen und Herausforderungen: Evaluation und Erfahrungen aus der Praxis*. Cornelsen Verlag.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E. (2009). The nature of generalization in language. *Cognitive Linguistics*, *20*(1), 93–127. https://doi.org/10.1515/COGL.2009.005

Goldberg, A. E., & Casenhiser, D. (2008). Construction learning and second language acquisition. In P. J. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 197–215). Routledge.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (Hrsg.). (2004). Learning argument structure generalizations. *Cogl*, *15*(3), 289–316. https://doi.org/10.1515/cogl.2004.011

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436. https://doi.org/10.1111/1467-9280.00476

Gómez, R. L., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, *7*(2), 183–206. https://doi.org/10.1207/s15327078in0702_4

Hempel, M. (2016). Förderung produktiver Sprachkompetenzen – Sind Lehrwerke Teil des Problems oder Teil der Lösung? In H. Böttger & N. Schlüter (Eds.), *Tagungsband zur 4. FFF-Konferenz 2014 in Leipzig* (pp. 124–133). Westermann.

Hopp, H. (2016). Learning (not) to predict: Grammatical gender processing in second language acquisition. *Second Language Research*, *32*(2), 277–307. https://doi.org/10.1177/0267658315624960

Hopp, H., & Thoma, D. (2021). Effects of plurilingual teaching on grammatical development in early foreign-language learning. *The Modern Language Journal*, *105*(2), 464–483. https://doi.org/10.1111/modl.12709

Ibbotson, P. (2013). The scope of usage-based theory. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00255

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. OUP.

Jaekel, N., Schurig, M., Schwinning, S., Ferencik-Lehmkuhl, D., & Ritter, M. (2024). Investigating the effects of linguistic distance on German and English reading and mathematics among fifth-graders in Germany. *Learning and Individual Differences*, 116. https://doi.org/10.1016/j.lindif.2024.102556

Kostka, N. (2020). *Produktives Sprechen im Englischunterricht der Grundschule—Eine empirische Studie zur Bedeutung formelhafter Sequenzen*. Universitätsbibliothek.

Kultusministerium Baden-Württemberg. (2004). *Bildungsplan 2004. Grundschule*. http://www.bildungsplaene-bw.de/site/bildungsplan/get/documents_E-1204433591/lsbw/Bildungsplaene/Bildungsplaene-2004/Bildungsstandards/Grundschule_Bildungsplan_Gesamt.pdf

Legutke, M., Müller-Hartmann, A., & Schocker- von Ditfurth, M. (2012). *Teaching English in the primary school*. Klett.

Li, F., & Sun, Y. (2024). Effects of different forms of explicit instruction on L2 development: A meta-analysis. *Foreign Language Annals*, *57*(1), 229–255. https://doi.org/10.1111/flan.12726

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching*, *46*(1), 1–40. https://doi.org/10.1017/S0261444812000365

Madlener-Charpentier, K. (2018). Do findings from artificial language learning generalize to second language classrooms? In A. E. Tyler, L. Ortega, M. Uno, & H. I. Park (Hrsg.), *Usage-inspired L2 Instruction: Researched pedagogy* (S. 211–234). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.49.10mad

Madlener-Charpentier, K. (2015). *Frequency effects in instructed second language acquisition*. De Gruyter.

Madlener-Charpentier, K., & Behrens, H. (2022). Konstruktion(en) erst- und zweitsprachlichen Wissens: Lernprozesse und Steuerungsoptionen aus gebrauchsbasierter Perspektive. In K. Madlener-Charpentier & G. Pagonis (Eds.), *Aufmerksamkeitslenkung und Bewusstmachung in der Sprachvermittlung: Kognitive und didaktische Perspektiven auf Deutsch als Erst-, Zweit- und Fremdsprache* (pp. 31–64). Narr Francke Attempto.

Madlener-Charpentier, K., & Pagonis, G. (2022). Optionen der didaktischen Formfokussierung zwischen Aufmerksamkeitslenkung und Bewusstmachung. In K. Madlener-Charpentier & G. Pagonis (Eds.), *Aufmerksamkeitslenkung und Bewusstmachung in der Sprachvermittlung: Kognitive und didaktische perspektiven auf Deutsch als Erst-, Zweit- und Fremdsprache* (pp. 11–30). Narr Francke Attempto.

McDonough, K., & Vleeschauwer, J. D. (2012). Prompt-type frequency, auditory pattern discrimination, and efl learners' production of wh-questions. *Studies in Second Language Acquisition*, *34*(3), 355–377. https://doi.org/10.1017/S0272263112000113

Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen (Ed.). (2012). *Richtlinien und Lehrpläne für die Grundschule in Nordrhein-Westfalen*. Ritterbach Verlag. https://www.schulentwicklung.nrw.de/lehrplaene/upload/klp_gs/LP_GS_2008.pdf

Nation, P., & Anthony, L. (2016). Measuring vocabulary size. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning*. *3* (pp. 355–368). Routledge.

Pagonis, G. (2022). Implizite Formfokussierung in der elementarpädagogischen Sprachförderung. In K. Madlener-Charpentier & G. Pagonis (Eds.), *Aufmerksamkeitslenkung und Bewusstmachung in der Sprachvermittlung: Kognitive und didaktische Perspektiven auf Deutsch als Erst-, Zweit- und Fremdsprache* (pp. 127–167). Narr Francke Attempto.

Pulido, M. F. (2024). Optimizing the input for learning of L2-specific constructions: The roles of Zipfian and balanced input, explicit rules and working memory. *Studies in Second Language Acquisition*, *46*(2), 379–403. https://doi.org/10.1017/S0272263124000081

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Ramscar, M. (2021). How children learn to communicate discriminatively. *Journal of Child Language*, *48*(5), 984–1022. https://doi.org/10.1017/S0305000921000544

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science 34*, 909–957. https://doi.org/10.1111/j.1551-6709.2009.01092.x

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences*, *26*(6), 462–483. https://doi.org/10.1016/j.tics.2022.03.007

Robinson, P. (2005). Cognitive abilities, chunk-strength, and frequency effects in implicit artificial grammar and incidental L2 learning: Replications of Reber, Walkenfeld, and Hernstadt (1991) and Knowlton and Squire (1996) and their relevance for SLA. *Studies in Second Language Acquisition*, *27*, 235–268. https://psycnet.apa.org/doi/10.1017/S0272263105050126

Robinson, P., & Ellis, N. (Eds.). (2008). *Handbook of cognitive linguistics and second language acquisition*. Routledge.

Schulz, J., Hamilton, C., Wonnacott, E., & Murphy, V. (2023). The impact of multi-word units in early foreign language learning and teaching contexts: A systematic review. *Review of Education*, *11*(2). https://doi.org/10.1002/rev3.3413

Silvey, C., Dienes, Z., & Wonnacott, E. (2025). Bayes factors for mixed effect models. *Psychological Methods*.

Simms, N. K., Frausel, R. R., & Richland, L. E. (2018). Working memory predicts children's analogical reasoning. *Journal of Experimental Child Psychology*, *166*, 160–177. https://doi.org/10.1016/j.jecp.2017.08.005

Twomey, K. E., Ranson, S. L., & Horst, J. S. (2014). That's more like it: Multiple exemplars facilitate word learning. *Infant and Child Development*, *23*(2), 105–122. https://doi.org/10.1002/icd.1824

Viviani, E., Ramscar, M., & Wonnacott, E. (2026). Go above and beyond: Does input variability affect children's ability to learn spatial adpositions in a novel language? *Royal Society Open Science*, *13*(2). https://doi.org/10.1098/rsos.251853

Wechsler, D. (1992). *Wechsler individual achievement test*. Psychological Corporation.

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence – Second Edition* (WASI-II) [Database record]. APA PsycTests. https://doi.org/10.1037/t15171-000

Wonnacott, E., Boyd, J. K., Thomson, J., & Goldberg, A. E. (2012). Input effects on the acquisition of a novel phrasal construction in 5 year olds. *Journal of Memory and Language*, *66*(3), 458–478. https://doi.org/10.1016/j.jml.2011.11.004

Year, J., & Gordon, P. (2009). Korean Speakers' Acquisition of the English Ditransitive Construction: The Role of Verb Prototype, Input Distribution, and Frequency. *The Modern Language Journal*, *93*(3), 399–417. https://doi.org/10.1111/j.1540-4781.2009.00898.x

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Accessible Summary

Appendix S1. Materials and Procedures for Matrix Reasoning Subtest of the Wechsler Abbreviated Scale of Intelligence for Children (WASI).

Appendix S2. Video Demonstrations of Each Approach Verb.

Appendix S3. Example Input Sentence Sets During Teaching.

Appendix S4. Sample Test Sentences for a Child in Each Exposure Condition.

Appendix S5. Pilot and Pilot Results.

Appendix S6. Detailed Testing Procedure During Outcome Measurement.

Appendix S7. Correlation Analyses Between Pretests and Outcome Measurement.

Appendix S8. Hypotheses to Be Tested and How They Relate to the Effects in the Different Models.

Appendix S9. Detailed Model Outcomes for Coding Type Linking Rule Correct (Without Verb Accuracy).

Appendix S10. Detailed Model Outcomes for Coding Type Entirely Correct (Including Verb Accuracy).
Appendix S11. Detailed Model Outcomes Forced-Choice Task.