Journal of Child Language (2018), 1–19
doi:10.1017/S0305000918000041

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

When regularization gets it wrong: children over-simplify language input only in production

Jessica F. SCHWAB*, Casey LEW-WILLIAMS, and Adele E. GOLDBERG

Department of Psychology, Peretsman-Scully Hall, Princeton University, Princeton, NJ 08540
*Corresponding author. Email: jschwab@princeton.edu

(Received 25 April 2017; revised 15 December 2017; accepted 8 January 2018)

Abstract

Children tend to regularize their productions when exposed to artificial languages, an advantageous response to unpredictable variation. But generalizations in natural languages are typically conditioned by factors that children ultimately learn. In two experiments, adult and six-year-old learners witnessed two novel classifiers, probabilistically conditioned by semantics. Whereas adults displayed high accuracy in their productions – applying the semantic criteria to familiar and novel items – children were oblivious to the semantic conditioning. Instead, children regularized their productions, over-relying on only one classifier. However, in a two-alternative forced-choice task, children's performance revealed greater respect for the system's complexity: they selected both classifiers equally, without bias toward one or the other, and displayed better accuracy on familiar items. Given that natural languages are conditioned by multiple factors that children successfully learn, we suggest that their tendency to simplify in production stems from retrieval difficulty when a complex system has not yet been fully learned.

Keywords: language acquisition; generalization; probability boosting

Introduction

Recent work has emphasized children's tendency to simplify (i.e., REGULARIZE) irregular or probabilistic input (Singleton & Newport, 2004; Hudson Kam & Newport, 2005, 2009). For example, a deaf child, Simon, was exposed to obligatory motion classifiers in American Sign Language (ASL) only 70% of the time by his parents, who were not native signers. Simon regularized his use of the classifiers to 90% of appropriate contexts, resulting in use that was indistinguishable from native signers (Singleton & Newport, 2004). Similarly, children who were taught a miniature artificial language that randomly included a classifier 60% of the time tended to regularize the pattern, using it more consistently than it had been witnessed. Adults, on the other hand, were more likely to replicate the probabilities witnessed in the input, producing the classifier approximately 60% of the time (Hudson Kam & Newport, 2005). As the input becomes more complex, however, adults also show a tendency to regularize (Hudson Kam & Newport, 2009; Harmon & Kapatsinski, 2017), particularly when adults are under pressure to communicate (Fehér, Wonnacott, & Smith, 2016). In each of these studies, the variation in the input was UNCONDITIONED, i.e., the choice between two or

more alternatives was random and therefore unpredictable. Because of this, regularization can be viewed as advantageous, as suggested by the title 'Getting it right by getting it wrong' (Hudson Kam and Newport, 2009): no predictable or meaningful distinctions are lost, and the language is made simpler and therefore easier to use.

In natural languages, however, unconditioned grammatical variation is vanishingly rare; instead, grammatical choices are typically conditioned by a combination of lexical, phonological, semantic, discourse, and sociolinguistic factors (Quirk, 1960; Bolinger, 1977; Trudgill, 2011). The fact that multiple factors are involved leads to generalizations that contain subregularities and exceptions. For example, English verb agreement is generally determined by the semantic number of the subject argument, but the word *pants* is treated as plural (*pants are*); sports teams are treated as singular in American English (*Manchester United is*), but optionally plural in British English (*Manchester United are*). Similarly, Spanish speakers learn that words ending in /o/ are generally grammatically masculine and so occur with the determiner *el*, while words ending in /a/ are generally feminine and occur with *la*, and they also learn lexical exceptions such as *el aroma* and *la mano*. Thus, in order to become a fluent speaker of a natural language, children must learn conditioned variation, including semantic generalizations that have lexical exceptions.

In work using miniature artificial languages that are reliably and exclusively CONDITIONED BY A SINGLE FACTOR, e.g., by lexical items, both adults and children have been found to reproduce the lexically conditioned variation. For example, if one word-order construction is witnessed with one set of verbs and a distinct word-order construction is witnessed with a different set of verbs, and if the two constructions are functionally equivalent so that the distinction is not conditioned by a difference in meaning or discourse context, adults display a strong tendency to reproduce each verb in whichever construction it had been witnessed (Wonnacott, Newport, & Tanenhaus, 2008; Perek & Goldberg, 2015; Thothathiri & Rattinger, 2016). Children, too, have been found to reproduce lexical conditioning in a similar paradigm in which two particles are reliably and exclusively conditioned by two different sets of nouns (Wonnacott, 2011).

As just noted, however, generalizations in natural languages typically involve multiple conditioning factors. This situation has been explored in previous work with adults. To simplify the discussion, we only consider input in which two conditioning factors exist in the input, and they are not witnessed conflicting with one another (cf. Perek & Goldberg, 2017, Exp. 2). For example, in one type of experiment, two distinct novel word-order constructions are conditioned by distinct discourse functions (Perek & Goldberg, 2015) or semantics (Thothathiri and Rattinger, 2016; Perek & Goldberg, 2017), as well as by lexical items. For example, in Perek and Goldberg (2017, Exp.1), adults learned six novel verbs, three of which exclusively appeared in one construction and three exclusively in the other. During exposure, one of the constructions was always used to describe scenes in which there was a weak effect on the undergoer argument, and the other construction was always used when there was a stronger effect on the undergoer argument. In a subsequent production task, some of the scenes displayed a weak effect on the undergoer but required a verb that had only been witnessed when the effect had been strong, or vice versa. Adults tended to use the verbs in whichever construction better conveyed the semantics of the scene, even if the verb had only been witnessed in the other construction. That is, when both lexical and semantic conditioning were equally valid cues to a construction, adults favored the semantic conditioning (see Perek &

Goldberg, 2015, for similar results regarding discourse conditioning). If the semantic conditioning was a MORE reliable cue than the lexical conditioning, because some of the verbs are witnessed occurring in both constructions, adults maximally applied the semantic conditioning, essentially ignoring the lexical conditioning (see also Thothathiri & Rattinger, 2016).

To summarize, if there is no functional reason to diverge from the input, adults aim to reproduce the variability in the input as it had been witnessed. If the input varies randomly by some proportion, they tend to reproduce variation in roughly the same proportion (Hudson Kam & Newport, 2005), and if it is lexically conditioned without attendant functional differences, adults reproduce the lexically conditioned input in their own productions (Wonnacott *et al.*, 2008; Perek & Goldberg, 2015; Thothathiri and Rattinger, 2016). But if the variation in the input is conditioned by discourse or semantic factors, adults readily generalize on that basis, as long as they have not witnessed lexical factors conflicting with the discourse or semantic conditioning in the input (Perek & Goldberg, 2015, 2017; Thothathiri & Rattinger, 2016).

There has been much less research that investigates children's behavior in miniature artificial language paradigms that include a combination of conditioning factors. Since, as noted, real languages are typically conditioned in just this way, it is important to explore how children's behavior compares to adults'. The present studies investigated what children and adults produce (Experiment 1) and what children select in a two-alternative forced-choice (2AFC) task (Experiment 2), when faced with a grammatical choice that is partially conditioned by a salient semantic cue: natural gender. Natural gender partially conditioned the input, in that a minority of cases were lexically conditioned (item-specific) because they referred to non-gendered (inanimate) entities. More specifically, each group of participants witnessed one novel classifier, e.g., *dax*, applied to three stereotypically female puppets (a mother, a girl, and a female dancer), and another novel classifier, e.g., *po*, applied to three stereotypically male puppets (a father, A boy, and a male doctor) (see Figure 1a). Two other inanimate puppets (a book and a ball) were lexically conditioned and assigned to one of the classifiers (*dax/po*) arbitrarily. That is, three out of four puppets paired with each classifier were predictable on the basis of natural gender, and two other puppets were lexically conditioned. Additionally, in order to investigate the possible role of regularization, one of the novel classifiers was witnessed twice as often as the other (e.g., the 'female' puppets were witnessed six times each, and the 'male' puppets only three times each; or vice versa). The asymmetric token frequency was motivated by the fact that previous work that had found a tendency to regularize had used asymmetric frequency distributions (Hudson Kam & Newport, 2005, 2009).

Experiment 1 tested whether six-year-old children and adults were equally able to reproduce the two classifiers when asked to describe familiar puppets, and whether they relied on the gender-based conditioning when referring to novel puppets. Experiment 2 probed a different group of six-year-olds on familiar and novel puppets in a 2AFC task.

Of particular interest was whether children and adults will be successful at appropriately producing the novel, semantically based classifiers with new gendered puppets (a grandma, a grandpa) and whether they will apply semantic or lexical conditioning when assigning classifiers to a MALE dancer and a FEMALE doctor. If participants use natural gender as a conditioning factor, we expect them to generalize on the basis of the perceived natural gender of the puppets, assigning the 'masculine' classifier to the male dancer and the grandpa puppet, and the 'feminine' classifier to

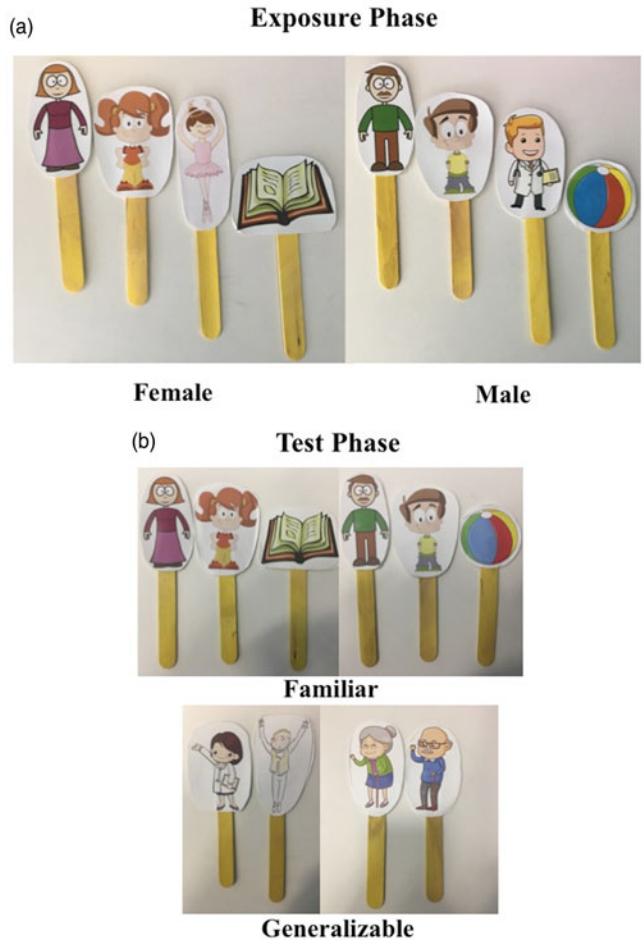

Figure 1. Puppets shown to participants during the Exposure Phase (a) and Test Phase (b).
Notes. The grouping of inanimate puppets (book/ball) into female/male categories was randomized across participants. During the Exposure Phase (a), one set of puppets (either 'Female' or 'Male') was witnessed twice as many times as the other set. During the Test Phase (b), each puppet was witnessed the same number of times.

the female doctor and the grandma puppet. If they instead rely on lexical conditioning, they should refer to the male dancer with the 'female' classifier, and the female doctor with the 'male' classifier, since they had witnessed those classifiers with the terms *doctor* and *dancer*, respectively, in the input; in this case they may be at chance at assigning classifiers to the grandma and grandpa puppets since these were not witnessed during exposure. Finally, if children tend to regularize complex input, even when a salient conditioning factor is available, they may regularize the entire system by boosting the probability of the more frequent classifier so that it applies categorically to all nouns. A second experiment tested a separate group of same-aged children on a 2AFC task after the same type and amount of exposure used in Experiment 1. The second experiment examined whether children PREFER formulations that obey a generalization, and if so, whether children's preferences align with their productions.

Previous work on conditioning that includes both semantic and lexical factors in adults led us to hypothesize that adults would readily detect the gender-based conditioning and would apply it productively to new items; that is, we expected adults to produce the 'masculine' classifier with both the grandpa and the male dancer puppets, and the 'feminine' classifier with the grandma and the female doctor puppets. The primary group of interest was children. Will children also generalize in a way that is consistent with the semantic distinction when required to describe new items? Will they behave in a lexically conservative manner by reproducing the lexical combinations that had been witnessed, without using the semantic conditioning to distinguish the classifiers? Or, as in previous studies when variation is unconditioned, will children regularize/simplify the input by over-relying on only one of the classifiers?

Experiment 1

Experiment 1 investigated adults' and children's productions of two novel classifiers that were probabilistically conditioned by the salient semantic feature of natural gender.

Method

Participants

Participants were 20 monolingual English-speaking children (M = 6;2; SD = 6.35; range = 5;5–7;1) and 20 monolingual English-speaking adults. Eleven children and 16 adults were female. Children were recruited through the Princeton Baby Lab and received a children's book and a small gift for participation, along with travel compensation for parents. Adults were recruited through Princeton University's Subject Pool, and participants received course credit. Children had no history of pervasive developmental delays. Six additional children were tested but not included due to fussiness/refusal to cooperate (n = 3), uninterpretable responses (n = 1), or instrument error (n = 2). Five additional adults were tested but not included in analyses due to being bilingual.

Stimuli and design

Participants were taught two different novel classifiers (*po* and *dax*). In 36 learning trials, each classifier was paired either with three male puppets (male doctor, father, and boy) or three female puppets (female dancer, mother, and girl), as well as one inanimate puppet (book or ball). In this way, the classifiers were probabilistically (75%) associated with natural gender. There was also a difference in classifier token frequency: 66.7% of practice trials used one classifier and 33.3% used the other. More specifically, four puppets occurred with the more frequent classifier (*dax* or *po*, counterbalanced) six times each, and the other four puppets occurred with the other (less frequent) classifier three times each.

Classifier/gender pairing, as well as the frequency of each classifier/gender pair in the input, was counterbalanced across participants. Learning trials consisted of a sentence describing a circular motion of each puppet, which was carried out by the experimenter. The form of the sentence was *moop* CLASSIFIER NOUN, where *moop* was a novel verb meaning 'moves in a circle', CLASSIFIER was one or the other classifier (*po* or *dax*) – probabilistically associated with either male or female gender – and NOUN was the English label for the puppet (e.g., 'boy'). Note that we describe the novel words *po* and *dax* as classifiers, but it is possible to construe them as agreement markers on the verb *moop*, since no morphemes intervened between the verb and

the 'classifiers'. Like classifiers, verbal agreement markers can be conditioned by natural gender cross-linguistically, so the interpretation of the present experiments does not hinge on this distinction.

During the test phase, five additional puppets were introduced: a female doctor, a male dancer, a grandma, a grandpa, and an apple (see Figure 1b).

Procedure

During the experiment, an experimenter sat across from each participant at a table. For children, the experimenter introduced a stuffed animal: "Mr. Chicken here talks funny! He says things differently from us. Today you're going to learn how to say some things the way that Mr. Chicken says them." Adult participants were instead told: "Today you're going to be hearing bits of a made-up language that uses puppets of familiar things in new ways. This study was designed for children, but we are interested in how adults learn language compared to kids."

Pre-exposure phase. The experimenter familiarized participants with 10 pictures that were identical to the cut-out puppets they would see during the exposure phase and/or test phase. The pictures were printed on construction paper and were displayed simultaneously. The experimenter pointed to one picture at a time and asked participants to help her figure out what names should be used for each picture in the language (for children: "What do you think Mr. Chicken calls this?", and for adults: "What do you think you call this, according to the language?"). Participants were given feedback and were retested on picture names until they provided the correct name for each picture.

Participants were then familiarized with each of the three novel words in the task (*moop, dax, po*), because they were going to be asked to produce these words themselves, and pilot testing revealed that children were often unable to remember the new words at test. The experimenter told participants: "You're going to be hearing 3 different new words from the made-up language, and I want to practice them together." Each participant was asked to practice saying *moop*, *dax*, and *po* six times each (for children, these were "words that Mr. Chicken uses" and for adults, these were "words from the made-up language"). All participants complied in practicing the words. Additionally, the order in which *dax* and *po* were introduced was counterbalanced across participants.

Exposure phase. At the beginning of the exposure phase, children were told: "We're going to keep learning how Mr. Chicken says things differently from us. I'm going to show you some things, and I'll say them how Mr. Chicken says them. Try to pay attention because after we practice together, you'll get a turn to try to say things the way Mr. Chicken does all on your own!" Adults were told: "Now you're going to learn how to say some things in the made-up language. I'm going to show you some things while saying a sentence in the language. For each sentence, I want you to repeat what I say. Try to pay attention because after we practice together, you'll get a turn to say things in the made-up language on your own." Children were also given a 'sticker board' (with 9 blank spaces, similar to a bingo board) and told they would get stickers along the way for helping the experimenter.

The experimenter pulled out puppets from a large box, one at a time. For each of the 36 learning trials, the experimenter would make a circular motion and say the accompanying sentence (in the form *moop* CLASSIFIER NOUN). After the first trial, the

experimenter would prompt the child to repeat the sentence while making the same circular motion (for children: "Now you try talking how Mr. Chicken talks", and for adults: "Now your turn"). In the following trials, the prompting sentence was omitted unless the participant was not responding. If the participant still did not say anything, the experimenter would prompt with the first word of the sentence, "Moop …".

Test phase. The test phase began with experimenter instructions (for children: "We're going to do a few more, but this time I won't be helping you. I'm going to show you some things, and I want you to tell me what you think Mr. Chicken would say," and for adults: "We're going to do a few more, but this time I won't be helping you. I'm going to show you some things, and I want you to say what you think would be an appropriate sentence using the made-up language"). Then the experimenter asked the participant if they remembered the three new words they had been practicing (*moop*, *dax*, and *po*). For the children, the experimenter had them rehearse all three new words two more times. For the adults, the experimenter only asked them to rehearse the three words if they said they did not remember them.

In each test trial, the experimenter again pulled out each puppet one at a time from the box and made a circular motion. Importantly, during test, the doctor and dancer puppets were gender-swapped compared to the exposure phase (e.g., puppets of a female doctor and male dancer were used in place of the male doctor and female dancer puppets used during the exposure phase), and novel grandma, grandpa, and apple puppets were added. Thus, there were six familiar test items (mother, father, girl, boy, ball, book), two gender-swapped test items (male dancer and female doctor), two novel test items that could be generalized based on gender (grandma and grandpa), and one neutral novel item (apple), for a total of 11 test items. Each test item was tested three times, for a total of 33 test trials, with the novel items (grandma, grandpa, apple) always tested last. The neutral test item (apple) was counted as a filler trial (because there is no 'correct' answer for its gendered classifier), so all following analyses include 30 test trials. Test orders were counterbalanced across participants (note that there were four test orders total, with two randomizations for the first 24 familiar-item test trials and two randomizations for the final nine novel-item trials). For each trial, if participants did not respond right away, the experimenter would again prompt with "Moop …". Both child and adult participants produced an identifiable classifier on every test trial.

Results and discussion

A comparison of adult and child performance accuracy is provided in Figure 2. As predicted, adults displayed ceiling level accuracy on the combinations of classifier + noun they had been exposed to ($M = .99$, $SE = 0.004$). While adults' accuracy was no doubt facilitated by their recognition of the semantic conditioning of gender, we note that they also performed at ceiling on the classifiers assigned arbitrarily to the two inanimate objects during exposure. Adults also took advantage of the semantic generalizations (male dancer, female doctor, grandma, grandpa) with high accuracy ($M = .90$, $SE = 0.04$).

In order to analyze the performance of the children, we fit a mixed effects logit regression to the data using the lme4 package in R (Bates, Maechler, Bolker, Walker, Christensen, & Singmann, 2015). We began with a maximal model that aimed to predict accuracy (in accord with natural gender and lexical conditioning for the

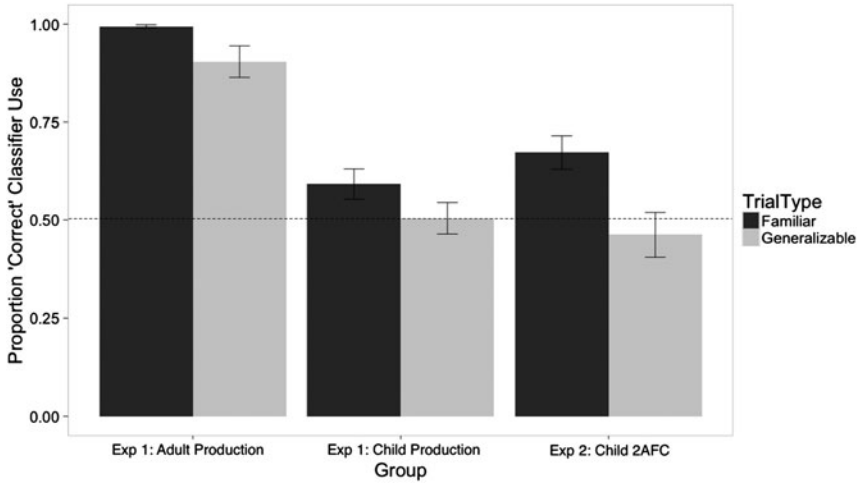

Figure 2. Proportion correct classifier use (i.e., accuracy) for participants in Experiment 1 (Adult and Child Production) and Experiment 2 (Child 2AFC) for familiar classifier/noun pairs and generalizable classifier/noun pairs.

Notes. Dashed line at 0.5 indicates chance performance. Error bars represent ±1 *SEM*.

inanimate puppets) from the fixed effect of Type of trial (familiar or generalizable), with random slopes and intercepts for subjects, and random intercepts for items, choice of high token frequency word (*po* or *dax*), and high token frequency gender (male or female). The last two random factors accounted for 0 variance, so were omitted from the model. We compared models using the ANOVA command, preferring a model with fewer degrees of freedom when the fit was a similarly good fit as a model with greater degrees of freedom. The same methodology was used in all models reported here, and raw data and analysis code can be accessed on Open Science Framework at <https://osf.io/z3eqa/>.

The final model predicted children's accuracy on the basis of the fixed effect of trial type (familiar or generalized), with random intercepts for subjects and items (and with the intercept term omitted to allow for chance comparison). Note that models with and without slopes for subjects were not significantly different (chi-square $p = .24$), so slopes were not included. Children's accuracy on the familiar items ($M = .59$, SE = 0.04) was better than chance ($\beta = 0.43$, $SE = 0.22$, $p = .046$), while their performance on novel puppets that required the semantic generalization (grandma, grandpa, female doctor, and male dancer) was not ($M = .50$, $SE = 0.04$) ($\beta = 0.04$, $SE = 0.25$, $p = .89$). Moreover, there was no indication that the children treated the novel female doctor and male dancer in a systematic way; only two out of the 20 children consistently used the masculine classifier with the doctor and the feminine classifier with the dancer, suggesting they may have been sensitive to the lexical conditioning in the input; however, even these two children failed to apply lexical conditioning systematically to the two inanimate entities provided in the input. The direct comparison between familiar and novel items (fixed effect of type, in a second model with intercept term included) was not significant ($\beta = -0.40$, $SE = 0.27$, $p = .14$).

Children's performance was far below the ceiling performance of adults (see Figure 2). This was confirmed with a mixed logistical regression that included fixed

effects of type (familiar or unfamiliar) and group (child or adult) and their interaction. Both fixed effects and their interaction were significant (all *p*s < .01) (note that including random intercepts for items or slopes for subjects resulted in non-convergent models).

Thus, children's production accuracy was low, particularly with combinations of classifier + noun that required generalization, where they performed at chance. When considered in aggregate, children were approximately twice as likely to produce the frequent classifier as the infrequent classifier (63% vs. 37%), which roughly reproduced the classifier frequencies in the input (67% vs. 33%). However, it would be misleading to infer that children were matching the input probabilities.

A closer look at individual behavior makes this clear (Figure 3). Adults selected each of the two classifiers roughly 50% of the time, as is required for high accuracy. On the other hand, children showed a marked tendency to simplify or regularize the input by over-relying on one classifier or the other. More specifically, we determined that participants would need to choose one of the classifiers on more than 69% or less than 31% of test trials in order to show behavior significantly different than chance (i.e., an underlying 50–50 distribution of classifier types) at a 95% confidence level, according to a binomial test (30 items/participant). In fact, 7 out of 20 children produced the more frequent classifier 100% of the time, 4 other children produced the more frequent classifier more than 69% of the time, and 6 children boosted the probability of the LESS frequent classifier more than 69% of the time, even though it had been witnessed only 33% of the time during exposure. Only 3 of 20 children regularly produced both classifiers (using each between 31% and 69%). We return to further analysis of this performance in a comparison with children's performance on the 2AFC task in Experiment 2.

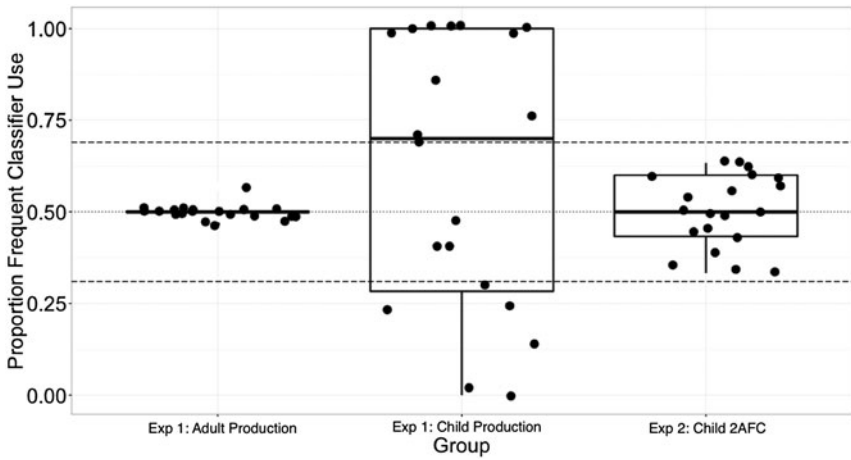

Figure 3. Box-and-whiskers plot showing proportion classifier use (based on input frequency in the exposure phase) for each participant's productions in Experiement 1 (Adult and Child Production groups) and for children's choices in Experiment 2 (Child 2AFC group).

Notes. Dotted line at 0.5 indicates equivalent frequent/infrequent classifier proportion use (in line with correctly learning the semantic generalization). Dashed lines at .31 and .69 indicate markers of probability maximizing for either the more frequent or less frequent classifier. Note that the majority of children fell outside this band in their productions (17 out of 20 children; Experiment 1), while no children fell outside this band in the 2AFC task (Experiment 2).

To be sure that children's failure to learn the gender-based distinction was not due to either (1) an inability to successfully use gender distinctions at all at this age or (2) an inability to correctly identify the gender of the puppets, we conducted a follow-up control study with a new group of six-year-olds. Participants were eight monolingual English-speaking children ($M = 6;0$; $SD = 5.63$; range = 5;6–6;7; none had a history of pervasive developmental delays; six were female). First, children were shown each of the 10 gendered puppets (i.e., all of the animate puppets) one at a time in a randomized order, and asked whether each puppet was a boy or girl. Next, children were asked to make up a sentence about each of the puppets. They were told they could say anything they wanted, but they should try to talk about the puppets in a similar way as the experimenter. The experimenter then gave example sentences using two NEW male and female puppets (e.g., "She has a blue skirt" and "He has a yellow shirt"). The children were then presented with each test puppet, one at a time again, in a randomized order, and asked to produce a description of each one. If children did not produce a full sentence (e.g., they simply said "yellow hair"), they were presented with the example sentences again. None of the children needed to hear the example sentences more than three times. All eight children successfully distinguished whether puppets were male or female 100% of the time (for all 10 puppets), and successfully used *he* or *she* appropriately with the puppets intended to be male and female, respectively, 100% of the time (for all 10 puppets). Thus, six-year-olds were unambiguously able to identify the genders of each of the puppets used in the task, and moreover, were able to successfully apply a gender distinction in selecting *he* or *she*.

To summarize the results of Experiment 1, our prediction for adults was confirmed: adults demonstrated recognition of the intended semantic distinction based on natural gender, and applied it to familiar and new gendered nouns with very high accuracy. In using whichever classifier was appropriate for the perceived gender of the puppets, adults overlooked the token frequency bias in the input and ignored the possible lexical conditioning in the case of the dancer and doctor puppets (see also Perek & Goldberg, 2015, 2017; Thothathiri & Rattinger, 2016). They also performed at ceiling on the two inanimate objects that were assigned classifiers arbitrarily.

Children's accuracy was markedly lower than that of adults. They were barely above chance on familiar items, and were at chance on items that required generalization. Instead, children displayed a tendency to regularize the input, disproportionately using one of the classifiers (usually, but not always, the more frequent classifier). Thus, children behaved as they do when faced with unconditioned variation between two options (Hudson Kam & Newport, 2005). However, in this case, since a salient reliable conditioning factor – natural gender – was available, it is not necessarily reasonable to say that children 'got it right' by regularizing the variable input.

Importantly, Experiment 1 tested children's PRODUCTION of novel classifiers. Given the potentially demanding nature of production tasks that require recalling multiple novel words, it is possible that children were simply relying on the most accessible classifier in memory. In fact, we know that accessibility exerts a strong influence on language production (e.g., MacDonald, 2013; Harmon & Kapatsinski, 2017). In the present context, producing one classifier increases the accessibility of that classifier (via repetition priming); therefore the easiest response is to produce the same classifier from one trial to the next. Without a solid grasp of the conditioning factors, there is no reason to move away from the easiest response (cf. inertia). Therefore, the tendency to perseverate on whichever classifier is more accessible might mask a tentative awareness that the target system is more complex.

In order to provide a different, potentially less demanding way for children to display an awareness of the conditioning factors in the input, Experiment 2 tested a second group of aged-matched children on a two-alternative forced-choice (2AFC) task, in which both classifier options were provided. This task allowed us to investigate whether children's over-reliance on one classifier in Experiment 1 implies a preference for consistent use of one classifier over the other, or whether children recognize that the use of the two classifiers was evenly distributed across the set of 12 puppets. (Recall that six puppets were assigned one classifier and six were assigned the other; only the token frequency was skewed, such that one set of puppets was witnessed twice as often as the other.) Since natural languages typically include variation that is conditioned by multiple factors, 'getting it right' would require children to ultimately learn the relevant conditioning factors. Is there any evidence that children are at least cognizant that the system included two distinct classifiers that apply to two distinct sets of entities?

Experiment 2

Method

Participants

Participants were 20 monolingual English-speaking children ($M = 6;1$; $SD = 6.68$; range = 5;5–7;2). Eight participants were female. Children had no history of pervasive developmental delays. Recruitment procedures were the same as in Experiment 1. Two additional participants were tested but not included due to instrument error ($n = 1$) or experiment error ($n = 1$).

Stimuli and design

The stimuli and design were based closely on the child version of Experiment 1. The experimenter's script differed slightly in order to replace the production task with a 2AFC task in the test phase.

Procedure

Pre-exposure and exposure phases. The pre-exposure and exposure phases were identical to the child version of Experiment 1, except that the experimenter instructed the child: "Try to pay attention because after we practice together, I'm going to ask you some questions about what Mr. Chicken says!"

Test phase. For the test phase, there was no additional practice with the three words (*moop, dax, po*), because the 2AFC task no longer had production demands. Instead, the test phase began with the following instructions: "Now I'm going to ask you some questions about what Mr. Chicken would say! But first, I have a question for you about what we would say." In order to give the child practice with the 2AFC format, the experimenter asked the child to judge a sentence in English. First, the experimenter held up a puppet of a car and asked the child: "What's this?" If the child responded appropriately (i.e., "car"), the experimenter would say: "That's right, that's a car." If the child did not respond, the experimenter would say "This is a car" and prompt the child to say "car" themselves. Next, the experimenter asked the child: "Now which one sounds better: 'I drive my car' or 'I swim my car'?" To clarify the break between the two options, the experimenter shifted her body slightly

for each one. This bodily shifting was done in order to provide children with a co-occurring visual cue to distinguish between the two response choices. If the child did not answer, the experimenter asked: "Drive or swim?" Once the child answered correctly, the experimenter proceeded to the main portion of the test phase (note that no children answered the practice question incorrectly).

The experimenter continued by saying: "Now I'm going to ask you some questions about what Mr. Chicken would say! Which do you think Mr. Chicken would say?" Next, on each of 30 test trials, the experimenter asked the child to answer whether Mr. Chicken would be more likely to say a sentence about each puppet that contained one or the other classifier (*dax* or *po*). For instance: "'Moop dax mother' or 'Moop po mother'?" Similar to the practice trial, for the first option, the experimenter shifted slightly to the left before making a circular motion with the puppet (to indicate *moop*). For the second option, the experimenter shifted slightly to the right before making a circular motion with the puppet. If the child did not answer right away, the experimenter would ask "dax or po?" (or "po or dax?" depending on the order of options previously given in that trial). Test orders were identical to Experiment 1 (in terms of order of puppets), and the ordering of *po* and *dax* options within trials was counterbalanced across participants. As in Experiment 1, all children chose an identifiable classifier on every test trial.

The main goal of Experiment 2 was to assess whether children were in fact unaware of the gender-based semantic distinction as indicated by Experiment 1, and if so, to determine whether they showed a PREFERENCE for utterances containing one classifier over the other.

Results

The proportion of familiar and unfamiliar combinations for which children chose the correct classifier is shown in Figure 2. A mixed linear regression model was created as described in the results section of Experiment 1, such that accuracy was predicted by the type of trial (familiar or generalizable) as a fixed effect, with random effects for subjects (intercepts and slopes) and items (intercepts). Note that in this model, including random slopes increases the fit of the model (chi-square $p = .005$), so slopes are included. An additional random intercept for whether the order of presentation on individual trials mattered was also included to account for any recency effect. In this model, children showed clear evidence of item-specific learning of the combinations of classifier + noun they had been exposed to ($M = .67$, $SE = 0.04$), performing significantly better on the familiar items compared to chance ($\beta = 0.83$, $SE = 0.26$, $p < .01$). However, in terms of classifier preference for the novel generalizable instances, children again failed to generalize according to natural gender when compared to chance ($M = .46$, $SE = 0.06$) ($\beta = -0.14$, $SE = 0.28$, $p = .60$), and they were also significantly worse on novel items compared to familiar items ($\beta = -0.97$, $SE = 0.23$, $p < .0001$).

To compare children's accuracy across the production task of Experiment 1 and the comprehension task of Experiment 2, we fit another mixed effect logit regression predicting accuracy as an interaction of trial type (familiar vs. generalizable) and experiment (production vs. 2AFC), with random effects of subject, high-frequency word (*po* or *dax*), and high-frequency gender (male or female). Type and experiment were sum coded. We again adopted the maximal model, simplifying the effects structure only if doing so resulted in a model that was a similarly good fit. This led to a model that included random intercepts for subjects and items. Results showed that

children across experiments performed better on familiar compared to generalizable trials, and that children's performance on familiar combinations was significantly better in Experiment 2 (2AFC task) than in Experiment 1 (Production task). That is, there was an effect of trial type such that children performed significantly worse on novel trials ($\beta = -0.66$, $SE = 0.19$, $p < .001$); there was also a significant trial type by experimental group interaction, demonstrating that children performed worse on familiar trials in the production task than in the 2AFC task ($\beta = -0.55$, $SE = 0.25$, $p = .03$) (see Table 1). We confirmed that the interaction added significantly to the model fit by comparing this model to one without the interaction (chi-square $p < .03$).

Of particular interest was the individual behavior in Experiment 2 compared with the production task of Experiment 1 (recall Figure 3). Children did not regularize in the 2AFC task (Experiment 2; rightmost panel) the way children did in the production task (Experiment 1; middle panel). Critically, not a single child in Experiment 2 showed a preference (above 69%) for either classifier, whereas in Experiment 1, 17/20 children produced classifiers in a way that was outside the 95% confidence intervals of what would be expected by chance. The difference in behavior between the two types of tasks was confirmed by a comparison of the absolute deviations of children's proportion use of the high-frequency classifier from chance in the two experiments. While children in Experiment 1 deviated from 50% use of the more frequent classifier in production on average ($|M| = .33$), children in Experiment 2 deviated from a 50% preference an average ($|M| = .09$), which was significantly less ($t(38) = 6.04$, $p < .001$, $d = 1.96$).

General discussion

While adults readily learned the semantic conditioning of the two classifiers, applying the distinction to familiar and novel gendered puppets alike, six-year-old children displayed no evidence of using natural gender to generalize in either the production task of Experiment 1 or the two-alternative forced-choice task of Experiment 2. Children did show some evidence of learning on the basis of lexical conditioning, particularly in Experiment 2. That is, children displayed better accuracy (of familiar

Table 1. Coefficient Estimates from a Generalized Linear Mixed Model (with a Logit Link) Predicting Children's Production or Choice of 'Correct' Classifiers

Child production vs. Child 2AFC task M1 and 2.kids <– glmer(accuracy ~ Type*Experiment + (1 |Subject) + (1 |Item), data = Exp1 and 2.kids, family = binomial)				
	Estimate	*SE*	*z*	*p*(|z|)
Intercept	0.3058	0.1525	2.005	**0.0449***
Trial type (Familiar)	−0.6563	0.1884	−3.483	**0.0005*****
Group1 (2AFC)	−0.0586	0.2116	−0.277	0.7818
Trial type × Group1	−0.5523	0.2528	−2.185	**0.0289***

Notes. Effects (Trial, Experiment) were sum coded (i.e., levels of each factor were compared to the average). Maximal random effect structure was simplified until models converged. This model compares six-year-olds' accuracy in production from Experiment 1 with accuracy in 2AFC in Experiment 2, with fixed effects of Trial type, Experiment (children from Experiment 1 and children from Experiment 2), and their interaction, along with random intercepts for subjects and items. Tests reaching statistical significance at the .05 criterion are marked in bold. *$p < .05$, **$p < .01$, ***$p < .001$.

classifier–noun combinations) in the 2AFC task than in the production task. Moreover, children showed no bias toward one classifier over the other in the 2AFC task, unlike children in the production task, who showed a strong tendency to over-rely on one classifier (thereby appearing to regularize).

Six-year-old children's insensitivity to the natural-gender distinction may seem surprising given that children at this age are well aware of the difference between boys and girls. In fact, six-year-olds performed at ceiling on the norming study that required them to identify the intended gender of each puppet and to use gendered pronouns (*he* and *she*) appropriately. However, consistent with the present results is work on the learning of gender marking in natural languages. When asked to assign gender to novel nouns, young speakers of Romance languages, aged three to twelve, have been found to overlook natural gender in favor of phonological cues (see Karmiloff-Smith, 1979; Pérez-Pereira, 1991; Surridge, 1993; Carroll, 2005). The reliance on phonology over semantics as a cue to gender marking may have multiple causes. In Romance languages, natural gender is a highly reliable cue to grammatical gender (words for 'boy' and 'girl' are masculine and feminine, respectively), but it is not a very available cue, since the vast majority of nouns refer to inanimate referents. In the present experiment, natural gender was much more available than in Romance languages (75% of the puppets were animate), but it was not uniformly available (25% were not). Recent work confirms that cues that are not uniformly available or reliable are more difficult for children to learn (Samara, Smith, Brown, & Wonnacott, 2017). Other recent work has found that cues that are available EARLIER in learning tend to be relied on more heavily than other cues even if the earlier cues are less salient (Culbertson, Gagliardi, & Smith, 2017). This may well play a role in children's relatively slow reliance on gender for classifiers in natural languages, since phonological cues are available to children from very early on (Gerken, Wilson, & Lewis, 2005). However, since the present experiment randomized the order of animate and inanimate trials during exposure, neither the lexical nor semantic conditioning factor was witnessed before the other in a systematic way.

Other work has found younger children to be less adept at discerning semantic conditioning cues when compared directly to older children and adults. For example, in a novel construction-learning paradigm in which constructions are assigned abstract meanings, five-year-olds have been found to display a much weaker ability to apply the appropriate generalization in new contexts when compared to seven-year-olds and adults (Boyd & Goldberg, 2009; Ferman & Kami, 2010; Raviv & Arnon, 2017). To summarize, young children are less adept at recognizing relevant conditioning factors, particularly when the factors are not uniformly available. Therefore, attending to probabilistically conditioned natural gender in the context of learning classifiers may be especially difficult for children.

Before the relevant conditioning factors are well-learned, children are likely to be unduly influenced by whichever option is more accessible at retrieval (MacDonald, 2013; Harmon & Kapatsinski, 2017). In the context of Experiment 1, we propose that the classifier retrieved on recent trials becomes more accessible for retrieval on future trials via repetition priming, resulting in a tendency to perseverate, or regularize. This is also suggested by Hudson Kam and Chang (2009, p. 816) who note: "when retrieval is difficult, the most easily accessible form is likely to be retrieved repeatedly, resulting in regularization." At the same time, others have argued that regularization is NOT the result of memory demands. Perfors (2012), in particular, has argued that regularization requires a "prior bias for regularization", on

the basis of experimental work showing that regularization did not increase when adults were put under cognitive load during exposure. However, as Perfors observed, that work investigated a possible role for memory ENCODING, whereas we are locating the issue as an effect at RETRIEVAL. Relevantly, retrieval is primarily an issue in production, and children tend to regularize in the production task of Experiment 1; retrieval is much less of an issue in 2AFC tasks, since both options are provided, and in fact, we find no tendency to regularize in that task (Experiment 2).

Children ultimately MUST learn complex sets of conditioning factors in order to select appropriate grammatical options in particular contexts. This brings us back to the question that motivated the second experiment: Do children tend to simplify the input in their productions because they are permanently converging on a more regular system, or because it is easier to retrieve the same choice due to repetition priming before conditioning factors have been well-learned? The results of Experiment 2 support the second interpretation. Clearly, the gender-based conditioning was not well-learned, as children were at chance on novel items in both experiments. In the 2AFC task, however, in which retrieval demands were mitigated, children showed no tendency to 'regularize' or over-rely on one classifier over the other. Instead, they displayed an awareness of two key aspects of the system: (a) that classifier–noun combinations were lexically conditioned, as they were above chance on familiar items; and (b) that each classifier was appropriate with roughly half of the nouns. They also displayed better accuracy on familiar items in the 2AFC task as compared to the production task. Performance on this task, therefore, undermines the idea that children are permanently acquiring a system that only involves one classifier and is therefore more regular.

In fact, the difference in children's behavior in production and 2AFC tasks extends to work in which the input includes unconditioned variation. Hudson Kam and Newport (2005) had also included both a production and a 2AFC task. While they emphasized that children tended to regularize the unconditioned input (as reviewed in the 'Introduction'), they also report that in the 2AFC task, children chose utterances that contained lower-frequency classifiers EQUALLY compared to utterances that contained the higher-frequency classifier (Figure 13 of Hudson Kam & Newport, 2005). While regularization was a reasonable response to unconditioned variation in that work, the fact that the children also regularized in the present Experiment 1, when the input contained systematic conditioning factors, suggests that children's tendency to 'regularize' by PRODUCING only one available option may mask a more accurate representation of the input that is revealed when both options are provided in 2AFC tasks.

The fact that children tend to regularize inconsistent input in production has been proposed as a mechanism by which languages become more regular over time (Hudson Kam & Newport, 2005; Reali & Griffiths, 2009). It is sometimes the case that learners do permanently regularize truly unconditioned input (e.g., when produced by non-proficient speakers) (Singleton & Newport, 2004). Speakers of every age prefer predictable over random variation (Kirby, Cornish, & Smith, 2008; Reali & Griffiths, 2009; Smith & Wonnacott, 2010). In other cases, learners imbue random variation with meaning by imposing conditioning factors in cases where none had existed (Janda, 1996; Smith & Wonnacott, 2010; Eckert, 2012; Smith, Perfors, Fehér, Samara, Swoboda, & Wonnacott, 2017).

It is worth noting that, even if children's regular productions were to provide the input to other learners, different children tend to regularize in different – in fact, opposing – ways. In aggregate, the children in Experiment 1 produced the two

classifiers with roughly the same probabilities that had been witnessed in the input; no regular system emerged from the combined productions across children. That is, the pooled production across children was no more regular than the initial input. The same is true of the children who witnessed unconditioned variation in Hudson Kam and Newport (2005): half of the children systematically used the single classifier that had been witnessed, and another 25% of the children systematically OMITTED the classifier in their own productions. Thus, there is a tendency to simplify in the direction of the more frequent option, but a minority of children in both Hudson Kam and Newport (2005) and in the present study simplified by boosting the probability of the less frequent option. Indeed, Smith *et al.* (2017) has found in modeling and experimental work that a more regular system does not tend to emerge, even in the iterated learning paradigm, when various speakers regularize in different ways.

The facts of natural languages require that if there is a systematic pattern to be learned, even if it is complex, children are very likely to succeed in learning it. In fact, sociolinguists have long argued that adolescents, not children, are the primary drivers of language change (Bybee & Slobin, 1982; Labov, 2001; Tagliamonte & D'Arcy, 2009; Trudgill, 2011; Evans Wagner & Tagliamonte, 2016).

To summarize, we wish to emphasize three points. First, children do not recognize relevant conditioning factors as readily as adults, even when the basis for the generalization is a salient semantic factor that is demonstrably familiar to them. Second, we suggest that children produce more regular (i.e., repetitive) utterances because repetition priming makes it easier to retrieve the same item again, and children tend to produce whatever is easier before the relevant conditioning factors have been well-learned. Third, in selecting between competing utterances, children display sensitivity to the complexities of the target system before they are capable of producing the complexity themselves.

Conclusion

Two mini-language learning studies demonstrate that six-year-old children are markedly less skilled at identifying a probabilistic semantic conditioning factor when compared with adults. Specifically, adults readily recognized and extended natural gender as a key conditioning factor for two novel classifiers, on the basis of the probabilistic association with natural gender in the input. However, six-year-old children revealed no awareness that gender was a relevant conditioning factor, despite the fact that at this age they are able to identify the intended genders of each puppet and are able to use the pronouns *he* and *she* accurately, as confirmed in a control study.

Instead of using natural gender as a conditioning factor, in the production task (Experiment 1), children showed a tendency to over-rely on one classifier or the other – usually, but not always, the more frequent classifier; their accuracy, even on familiar items, was only just above chance. Thus, six-year-old children exposed to variation in the input that was conditioned probabilistically behaved much like children in previous work who were faced with unconditioned variation (Hudson Kam & Newport, 2005). In both cases, six-year-olds regularized the input in the sense that they tended to boost the probability of one classifier. Thus, when children witness input that they perceive to be unconditioned, either because it IS unconditioned (Hudson Kam & Newport, 2005) or because they fail to recognize the conditioning factor (as in the present study), children tend to simplify the language in their own productions.

At the same time, when accessibility demands are reduced because both options are provided in a two-alternative forced-choice task (Experiment 2), children reveal better item-specific learning, and display no preference for one classifier over the other. The recognition that both classifiers are equally appropriate, albeit with different words, would allow memory of specific classifier + noun combinations to accrue over time, ultimately providing children access to the more complex system which involves both semantic and lexical conditioning factors. A similar difference in behavior in production and 2AFC tasks has also been found when variation in the input is unconditioned (Hudson Kam & Newport, 2005), which raises questions about the interpretation of regularization in production in that case as well. Since natural languages disprefer random variation, two possible long-term outcomes are possible. The first is that only one option continues to be used (a solution that the deaf child, Simon, converged on in his use of classifiers in ASL). Another option is that learners imbue the random variation with meaning, often sociolinguistic, such that conditioning factors emerge over time (Janda, 1996; Smith & Wonnacott, 2010; Eckert, 2012; Smith *et al.*, 2017).

The present work suggests that when conditioning factors exist in language input, as is almost always the case when learning from native speakers, children's tendency to regularize variable input in their productions does not represent the end state of their learning. Rather, the regularization appears to arise from the fact that one option is more accessible during retrieval. That is, during the period when children's learning is tentative, there is little motivation to move away from the easiest option, and this leads children to perseverate on that option. Since this 'regularization' occurred in Experiment 1, even though the input was conditioned by salient semantic and lexical factors, and since children displayed better accuracy and no tendency to regularize when accessibility demands were reduced (in Experiment 2), we suggest that children's tendency to regularize in production provides only a temporary snapshot of their language learning from conditioned input. Perhaps the strongest argument for this position comes from natural languages themselves. If children were to permanently regularize probabilistically conditioned input, natural languages would be much more regular than they actually are. Instead, generalizations in natural languages are typically conditioned by a variety of factors that are less than 100% reliable or available.

Acknowledgements. This research was funded by a grant from the National Institute of Child Health and Human Development to CLW (R03HD079779) and a Cognitive Science Graduate Fellowship from Princeton University to JFS. We thank Jessica Quinter for help with data collection, as well as Eva Fourakis and the rest of the Princeton Baby Lab for administrative assistance. We are grateful to Kenny Smith and an anonymous reviewer as well as to Associate Editor Caroline Rowland for thoughtful and helpful feedback on an earlier version of this paper. We also thank Ting Qian for statistical advice.

References

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., & Singmann, H. (2015). lme4: linear mixed-effects models using Eigen and S4 [Computer software manual]. Retrieved from <http://CRAN.R-project.org/package=lme4> (R package version 1.1-7).

Bolinger, D. L. M. (1977). *Meaning and form*. London: Longman.

Boyd, J. K., & Goldberg, A. E. (2009). Input effects within a constructionist framework. *Modern Language Journal*, *93*, 418–29.

Bybee, J., & Slobin, D. (1982). Why small children cannot change language on their own: evidence from the English past tense. In A. Alqvist (Ed.), *Papers from the 5th International Conference on Historical Linguistics* (pp. 29–37). Amsterdam: John Benjamins.

Carroll, S. E. (2005). Input and SLA: adults' sensitivity to different sorts of cues to French gender. *Language Learning*, *55*, 79–138.

Culbertson, J., Gagliardi, A., & Smith, K. (2017). Competition between phonological and semantic cues in noun class learning. *Journal of Memory and Language*, *92*, 343–58.

Eckert, P. (2012). Three waves of variation study: the emergence of meaning in the study of variation. *Annual Review of Anthropology*, *41*, 87–100.

Evans Wagner, S., & Tagliamonte, S. A. (2016). *Vernacular stability: comparative evidence from two lifespan studies.* Paper presented at New Ways of Analyzing Variation (NWAV) 45, Vancouver, Canada.

Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language*, *91*, 158–80.

Ferman, S., & Kami, A. (2010). No childhood advantage in the acquisition of skill in using an artificial language rule. *PloS one*, *5*, e12648.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*, 249–68.

Harmon, Z., & Kapatsinski, V. (2017). Putting old tools to novel uses: the role of form accessibility in semantic extension. *Cognitive Psychology*, *98*, 22–44.

Hudson Kam, C. L., & Chang, A. (2009). Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 815–21.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*, 151–95.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: when learners change languages. *Cognitive Psychology*, *59*, 30–66.

Janda, L. A. (1996). *Back from the brink: a study of how relic forms in languages serve as source material for analogical extension*. Munich & Newcastle: Lincom Europa.

Karmiloff-Smith, A. (1979). Micro-and macrodevelopmental changes in language acquisition and other representational systems. *Cognitive Science*, *3*, 91–118.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*, 10681–5.

Labov, W. (2001). *Principles of linguistic change, vol. 2: social factors*. Oxford: Blackwell.

Macdonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*, 226.

Perek, F., & Goldberg, A. E. (2015). Generalizing beyond the input: the functions of the constructions matter. *Journal of Memory and Language*, *84*, 108–27.

Perek, F., & Goldberg, A. E. (2017). Linguistic generalization on the basis of function and constraints on the basis of statistical preemption. *Cognition*, *168*, 276–93.

Pérez-Pereira, M. (1991). The acquisition of gender: what Spanish children tell us. *Journal of Child Language*, *18*, 571–90.

Perfors, A. (2012). When do memory limitations lead to regularization? An experimental and computational investigation. *Journal of Memory and Language*, *67*, 486–506.

Quirk, R. (1960). Towards a description of English usage. *Transactions of the Philological Society*, *59*, 40–61.

Raviv, L., & Arnon, I. (2017). *Differences between children and adults in the emergence of linguistic structure.* Paper presented at the CUNY Conference on Human Sentence Processing.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: relating regularization to inductive biases through iterated learning. *Cognition*, *111*, 317–28.

Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: children and adults are sensitive to socially conditioned linguistic variation. *Cognitive Psychology*, *94*, 85–114.

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: the acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, *49*, 370–407.

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use, and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B*, *372*, 20160051.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*, 444–9.

Surridge, M. E. (1993). Gender assignment in French: the hierarchy of rules and the chronology of acquisition. *IRAL-International Review of Applied Linguistics in Language Teaching*, *31*, 77–96.

Tagliamonte, S. A., & D'Arcy, A. (2009). Peaks beyond phonology: adolescence, incrementation, and language change. *Language*, *85*, 58–108.

Thothathiri, M., & Rattinger, M. G. (2016). Acquiring and producing sentences: whether learners use verb-specific or verb-general information depends on cue validity. *Frontiers in Psychology*, *7*.

Trudgill, P. (2011). *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford University Press.

Wonnacott, E. (2011). Balancing generalization and lexical conservatism: an artificial language study with child learners. *Journal of Memory and Language*, *65*, 1–14.

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: distributional learning in a miniature language. *Cognitive Psychology*, *56*, 165–209.

Cite this article: Schwab JF, Lew-Williams C, Goldberg AE. When regularization gets it wrong: children over-simplify language input only in production. *Journal of Child Language* https://doi.org/10.1017/S0305000918000041