Uninversion error in English-speaking children's *wh*-questions: Blame it on the bigrams?

Ben Ambridge[1,2]
Stewart M. McCauley[3]
Colin Bannard[4,2]
Michelle Davis[1,2]
Thea Cameron-Faulkner[4,2]
Alison Gummery[2,5]
Anna Theakston[1,2]

University of Manchester, Division of Psychology, Communication and Human Neuroscience[1]
ESRC International Centre for Language and Communicative Development (LuCiD)[2]
University of Iowa, Department of Communication Sciences and Disorders [3]
University of Manchester, Linguistics and English Language[4]
University of Liverpool, Psychology[5]

Abstract: The aim of the present study was to investigate whether and how English-speaking children's uninversion errors with wh-questions (e.g., **Who he can draw; c.f., Who can he draw?*) are influenced by the surface frequency of individual bigrams and trigrams in the input, as predicted by input-based approaches. Production methods were used to elicit nonsubject *wh*-questions from 67 children aged 3;1 to 4;8 (*M*=4;0, *SD*=4 months). No support was found for the preregistered prediction that children will produce more uninversion errors when those errors incorporate – in the Bigram 3 position – high-frequency bigrams from uninverted structures (e.g., **Who he* ***can draw****?*) than lower-frequency bigrams from uninverted structures (e.g., **Who he* ***can name?***). Importantly, when testing this prediction, all other bigrams and unigrams (i.e., single words) are either identical (e.g., *Who+he, he+can, he, can*) or closely matched for frequency (e.g., *draw* and *name* [as verbs] are of approximately equal corpus frequency). However, a non-preregistered exploratory analysis found a facilitatory effect on correct-question production of the frequency of the second and third bigrams from inverted structures (e.g., *can he…he draw*), even after controlling for unigram frequency. This analysis also found that rates of uninversion error (e.g., **Who he can draw?*) were higher when the first uninverted bigram (e.g., *Who he…*) is of higher frequency in the input. We conclude that while input-based accounts are correct to highlight the importance of n-gram input frequencies on rates of correct production versus uninversion error, it is unclear on current evidence which n-grams are driving errors and why. In particular, the special emphasis placed by some such accounts on n-grams at the left-edge of the utterance (e.g, *Who can…*) may be unwarranted.

Keywords: *wh*-questions, elicited production, elicited imitation, frequency.

Corresponding author(s): Ben Ambridge, Psychology, Communication and Human Neuroscience, Coupland Building 1, University of Manchester, Manchester, UK, M15 6FH.

ORCID ID(s): https://orcid.org/0000-0003-2389-8477

Citation: Ambridge, B., McCauley, S., Bannard, C., Davis, M., Cameron-Faulkner, C., Gummery, A., Theakston, A. Uninversion error in English-speaking children's *wh*-questions: Blame it on the bigrams?. *Language Development Research, 3*(1), 121–155. https://doi.org/10.34842/2023.641

Introduction

Wh-questions occupy a special place in language development research, since they are the only commonly used sentence-level construction for which English-speaking children regularly produce word-order errors; specifically, uninversion (or non-inversion) errors[1] such as **Who he can draw?* (cf., *Who can he draw?*). Early interest in these errors (Bellugi, 1971; Hurford, 1975; Kuczaj, 1976; Tyack & Ingram, 1977; Maratsos & Kuczaj, 1978; Labov & Labov, 1978; Kuczaj & Brannick, 1979; Bloom, Merkin & Wooten, 1982; Erreich, 1984) was sparked by the fact that they appear to reflect children's failure to apply a particular form of syntactic movement (I-to-C movement, or *subject-auxiliary inversion*; e.g., *Who he can draw?* → *Who can he draw?*) having already moved the *wh-* word from its corresponding position in declarative utterances (e.g., *He can draw who* → *Who he can draw*).

Subsequent accounts developed in this movement- or rule-based framework have sought to explain why children fail to apply this movement rule to particular *wh*-words (DeVilliers, 1991; Valian, Lasser & Mandelbaum, 1992; Pozzan & Valian, 2017), auxiliaries (Santelmann, Berk, Austin, Somashekar, & Lust, 2002; Hattori, 2003; Westergaard, 2009), or both (Stromswold, 1990, 1995; Valian & Casey, 2003).

In contrast, accounts developed in a usage-based (or "constructivist") framework have sought to explain these errors (sometimes referred to as "non-target-consistent" or simply "ungrammatical" questions) in terms of properties of the input. We term these accounts "input-based" because – although all accounts must of course posit *some* role for the input – such accounts claim that children are learning the structure of questions directly from the input, rather than merely using the input to trigger rules or parameters (e.g., wh-movement; I-to-C movement' subject-auxiliary inversion). That said, as we will see shortly, different varieties of input-based account potentially make subtly different predictions regarding frequency effects in question production.

Consistent with input-based approaches (in the broad sense), several studies have shown that children are less likely to produce uninversion errors (e.g., **Who he can draw?*) when lexical strings that appear in the correct form – particular *wh*-word+auxiliary combinations, such as *who can* – are frequent in the input (e.g., Rowland & Pine, 2000; Rowland, Pine, Lieven & Theakston, 2003, 2005; Dabrowka & Lieven, 2005; Ambridge, Rowland, Theakston & Tomasello, 2006; Rowland, 2007; Ambridge & Rowland, 2009). McCauley, Bannard, Theakston, Davis, Cameron-Faulkner and Ambridge (2021) also showed that children are more likely to produce uninversion errors (e.g., **Who he can draw*) when lexical strings that appear in the errorful form are frequent in the input (e.g., *he can* is considerably more frequent than *can he*). Of course, these findings do not demonstrate the *absence* of a syntactic subject-auxiliary inversion rule. What they do suggest however, is that – at the very least – the output of such a rule is filtered through a production mechanism that is sensitive to the input frequency of multiword strings (e.g., bigrams and trigrams; collectively n-grams); a

[1] Technically, "uninversion" incorrectly implies that the erroneous questions started out as inverted, and were then "uninverted". However, because the term is more widespread in the literature than the slightly more cumbersome term "non-inversion" errors, we use it (and "uninverted") throughout.

mechanism that can both cause errors and protect against them (Ambridge, Rowland, Theakston & Kidd, 2015).

The aim of the present study was to conduct a particularly tightly controlled investigation of input-based accounts of uninversion errors by investigating the effect of the input frequency of the third bigram in uninverted questions, while holding constant the frequency of all other bigrams (e.g., **Who he **can draw?*** [high-frequency] vs. **Who he **can name?*** [low-frequency]). This constitutes something of a departure from most studies in this domain (McCauley et al., 2021, excepted), which have generally focused on n-grams towards the left edge of the utterance and – in the main – on n-grams that appear solely or mainly in questions (e.g., *who can* or *who can he*), and that therefore support correct-question formation, rather than causing uninversion errors. Having conducted a preregistered test of this prediction, we then go on to conduct exploratory analyses in which we investigate in a more open-ended fashion input-frequency effects for other n-grams; again, both n-grams from inverted structures (mainly questions) that protect against inversion errors, and n-grams from uninverted structures (mainly declaratives) that cause inversion errors.

The starting point for the present study is the corpus study of McCauley et al. (2021) which, in turn, was inspired by studies showing faster processing and/or fewer production errors for higher frequency n-grams, for both adults (e.g., Liberman, 1963; Krug, 1998; Bybee & Scheibman, 1999; Jurafsky, Bell, Gregory, & Raymond, 2001; Sosa & MacFarlane, 2002; McDonald & Shillcock, 2003; Pluymaekers, Ernestus, & Baayen, 2005; Bannard, 2006; Arnon & Snider, 2010; Tremblay & Baayen, 2010; Siyanova-Chanturia, Conklin, and van Heuven, 2011; Janssen & Barber, 2012; Hernández, Costa & Arnon, 2016; Arnon, McCauley & Christiansen, 2017) and children (Bannard & Matthews, 2008; Arnon & Clark, 2011; Havron & Arnon, 2021; Skarabela, Ota, O'Connor & Arnon, 2021; Kueser & Leonard, 2020).

In an analysis of 12 spontaneous speech corpora from the English-speaking portion of CHILDES (MacWhinney, 2000), McCauley et al. (2021) showed that the frequency of children's uninversion errors versus correct questions (e.g., **What you are doing there* vs *What are you doing there?*) was (a) negatively related to the input frequency of the third and fourth bigram in the correct, inverted question (e.g., *you doing; doing there*) and (b) positively related to the input frequency of the second, third and fourth bigram in the errorful, uninverted question (e.g., *you are, are doing, doing there*). To clarify, the reason that children were hearing "uninverted" bigrams such as *you are, are doing* and *doing there* was NOT because their caregivers were producing uninversion errors; they were not. Rather, children were hearing these "uninverted" bigrams as part of declarative sentences (e.g., ***You are** happy; They **are doing** it*), complement clauses (*I wonder what he's **doing there***), including those used for reported questions (e.g., *He asked whether **you are doing** it*), and so on. That is, even though these children were easily capable of distinguishing questions from declaratives and other non-questions, high-frequency uninverted n-grams heard in the context of declaratives constituted "lures" towards uninversion errors in question production; albeit lures that children could resist when the target inverted n-grams (i.e., those heard in the context of questions) were of sufficiently high input frequency. These findings are summarized in Figure 1 (reproduced from McCauley et al., 2021, under the terms of the Creative Commons CC-BY license, which permits unrestricted use).

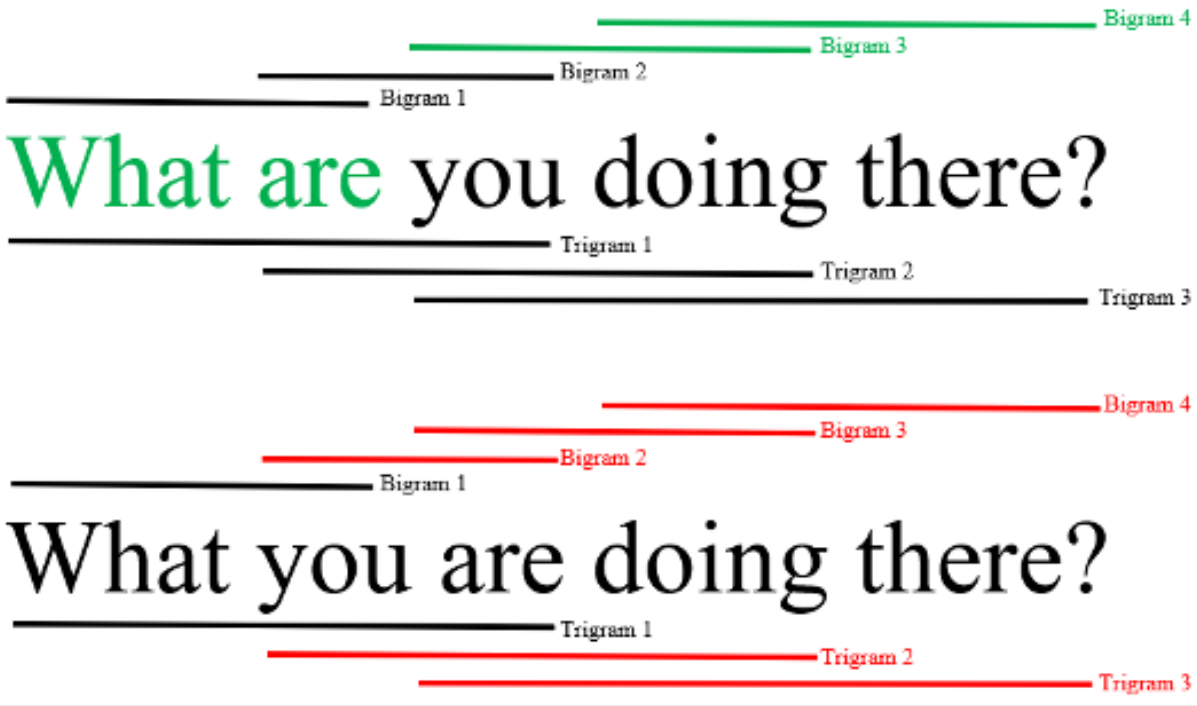

Figure 1. ***Summary of the findings of the corpus study of McCauley et al. (2021). Unigrams (individual words), bigrams, and trigrams for the correct, inverted (top) and corresponding errorful, uninverted (bottom) forms of the example question What are you doing there? N-grams excluded from the final statistical model are shown in black. N-grams retained in the final statistical model are shown as green/red words (unigrams) and green/red lines (bigrams and trigrams).***

Indeed, there is precedent for McCauley et al's (2021) finding that high frequency input strings from one sentence type (here, mainly declaratives) can constitute "lures" towards errors for a different sentence type (here, questions). For example, in Norwegian (like many V2 languages), the negation marker appears after the verb in main clauses (e.g., *We* ***read not*** *Icelandic sagas every night*) but before the verb in embedded clauses (e.g., *The teacher knows that we* ***not read*** *Icelandic sagas every night*). Children learning Norwegian often make errors when attempting to produce embedded clauses (e.g., **The teacher knows that we* ***read not*** *Icelandic sagas every night*), by inappropriately generalizing on the basis of high-frequency combinations with main-clause word order, here ***read+not*** (Westergaard & Bentzen, 2007; Ringstad & Kush, 2021; see also Waldmann, 2012, for a similar finding in Swedish). This is analogous to McCauley et al's (2021) finding that high-frequency n-grams from (mainly) declaratives (e.g., ***you are***) lead to uninversion errors in question formation (e.g., **What* ***you are*** *doing?*).

Perhaps surprisingly, unlike previous studies (e.g., Rowland & Pine, 2000; Ambridge et al., 2006; Ambridge & Rowland, 2009) McCauley et al. (2021) found no significant frequency effect of the first inverted bigram, which – for *wh*-questions – is always a *wh*-word+auxiliary combination (e.g., *What are*; *What is*; *Why is*; *Who can* etc…). However, this may be a consequence of the unusually strict analysis used by McCauley et al. (2021), under which bigram frequency effects were investigated only after controlling for frequency effects at the level of each individual lexical item (or "unigram").

Indeed, significant unigram frequency effects were observed for the first two inverted positions (e.g., *What; are*). Thus, we cannot conclude that the frequency of the first inverted bigram (*wh*-word+auxiliary) has no effect; only that we cannot detect an effect of the *wh*-word+auxiliary combination above and beyond frequency effects observed for the *wh*-word and auxiliary individually.

The aim of the present study was to conduct an experimental test of a prediction that follows from the study of McCauley et al. (2021), and from the more general claim of (at least some) input-based approaches, that learners retain, and are influenced by, individual lexical strings even when they have formed more abstract representations too (e.g., Langacker, 1998; Abbot-Smith & Tomasello, 2006; Goldberg, 2006; Ambridge 2020a, 2020b). That prediction, as preregistered at https://osf.io/tbmu4 prior to data collection (registration DOI: https://doi.org/10.17605/OSF.IO/TBMU4), was as follows:

> *Wh-* questions are more likely to be produced without subject-auxiliary inversion when the multiword sequences making up the errorful, non-inverted form of a child's target question are of a higher frequency. That is, subjects will make more uninversion errors with questions in the high frequency condition (where the frequency of "can go" in the uninverted form of the question "where can he go?" is high) than in the low frequency condition (where "can play" in the uninverted form of the question "where can he play?" is of a lower frequency, relative to "can go," while the correctly form[ed] questions are matched for the frequency of all trigrams, bigrams, and unigrams).

In order to have control over the target questions that children were attempting to produce, it was necessary to use an elicited-production methodology, in which the experimenter produced the target *wh*-word, auxiliary, subject and verb, but in uninverted order (as per Ambridge et al., 2006, 2008; Ambridge & Rowland, 2009). The method can be summarized as follows (again, quoting from our preregistration document):

> The experiment is couched in terms of a "jigsaw puzzle" game where the child is asking questions to a toy dog…In each trial, the child is prompted to produce a question by the experimenter by showing them an image consisting of one or more "jigsaw puzzle" pieces. Slots for missing jigsaw pieces are apparent in this image, and conceal some aspect of the target question. For instance, the missing jigsaw pieces may be hiding a ball in the case of a trial involving the target question "What is she holding?" The experimenter then attempts to elicit the target question by saying "I wonder what she's holding? Let's ask the dog what she is holding!" When the child asks the question, the missing jigsaw pieces are then filled in to reveal the ball (in the case of this example trial). The child then hears an audio recording (meant to be the dog's voice) answering the question. In this case, "A ball!"

Before setting out the present study in detail, it is important to clarify that not all "input-based" accounts of question acquisition would necessarily share the prediction set out above (or the non-preregistered effects that we uncovered in subsequent, exploratory analyses). For example, Rowland and Pine (2000), Dabrowska and Lieven

(2005), Ambridge et al. (2006) and Ambridge & Rowland (2009) all posit that children, certainly by age 3-4, form slot-and-frame question schemas such as *What are [THING] [PROCESS]?* Because these are informal, verbal accounts (as opposed to formal mathematical or computational models) they do not yield precise quantitative predictions. But one possible interpretation of these accounts – and quite possibly the dominant one in the literature – is that only the "frame" (e.g., *What+are*) is fixed, with the "slots" [THING] [PROCESS] free. Consequently, such accounts arguably predict that the frequency of words or combinations in the slot positions will not affect rates of correct production versus uninversion error.

In the present study, however, we test a different, more radically-exemplar-based type of input-based account (e.g., Langacker, 1998; Abbot-Smith & Tomasello, 2006; Goldberg, 2006; Ambridge 2020a, 2020b; McCauley et al., 2021).,which assumes that whether or not children form, in some sense, "free slots", they remain sensitive to the frequency of the individual n-gram combinations in the exemplars that gave rise to those slots.

Method

Ethics

Ethics approval was obtained from the University of Liverpool Research Ethics Committee prior to recruitment. Children's caregivers gave informed written consent and children gave verbal consent.

Participants

Our preregistration specified a minimum of 60 (providing 90% power) and a maximum of 70 participants, chosen on the basis of a power analysis calculation conducted using the "simr" R package (for details see https://osf.io/74urw/) assuming α=0.05. The simulation data were based on a small pilot study (N=12), but were adjusted to assume a small effect size for our primary manipulation (d=0.2), since no such effect was present in the pilot data. All children were native learners of English, with no known language impairments, and received stickers for their participation.

Given that our primary manipulation compares rates of uninversion errors within matched question pairs (e.g., *Who can he draw?* vs *Who can he name?*), it was important to ensure that we recruited a sufficient number of participants who produced scoreable responses (correct questions or uninversion errors with the target lexical items) for *both* questions in a given pair. Our preregistration therefore stipulated that "We will retain data only from children who produce scorable responses (correct question or noninversion errors) for a minimum of three high+low frequency pairs. Any excluded participants will be replaced in order to ensure our target sample size of 60". Of the 113 children who began the study, 46 were excluded and replaced on this basis, for a final sample size of N=67. Although a drop-out rate of 40% may seem high, it partly reflects the fact that – due to our focus on particular n-grams – it was necessary to exclude otherwise-scorable questions that included perfectly reasonable substitutions (e.g., *Who can the man draw?* for *Who can he draw?*). The final sample ranged in age from 3;1 to 4;8 with a mean of 4;0 (SD=4 months).

Design and Materials

The primary aim of the study was to test the prediction that participants will produce more uninversion errors when those errors incorporate – in the Bigram 3 position – high-frequency (e.g., **Who he **can draw**?*) rather than lower-frequency bigrams from uninverted structures (e.g., *Who he **can name**?*). Recall that when testing this prediction, all other bigrams and unigrams (i.e., single words) are either identical (e.g., *Who+he, he+can, he, can*) or closely matched for frequency (e.g., *draw* and *name* [as verbs] are of approximately equal corpus frequency). Using n-gram frequencies from the child-directed portion of the entire CHILDES database for both UK and US English (MacWhinney, 2000) – a total of 3,436,333 utterances (not selected or coded for sentence type) – we created eight question pairs that met this criterion (see Table 1). It is important to make clear at this point that, as explained in further detail below, the experimenter's prompt sentences in fact included uninverted questions; albeit grammatically acceptable ones that constitute reported speech (e.g., *I wonder who he can draw*). Thus in order to produce a well-formed question, the child has to "invert" the experimenter's prompt question.

With regard to the frequency of the n-grams from inverted structures, the "high" and "low" frequency questions in each pair (defined with regard to Uninverted Bigram 3) were perfectly matched for Bigram 1 and Bigram 2 (since the first three words were identical) and approximately matched for Bigram 3 (and likewise for the corresponding trigrams)[2]. For example, consider the question pair *Who can he draw? / Who can he name?*, which yield the uninversion errors **Who he can draw? / *Who he can name?*. With regard to the frequency of the n-grams from inverted structures, the two are identical with respect to Bigram 1 (*who can*) and Bigram 2 (*can he*), and closely matched for Bigram 3 (*he draw* = 15 occurrences in the corpus; *he name* = 12 occurrences). The high and low frequency questions of each pair were closely matched with regard to the frequency of n-grams from inverted structures, in order to allow for a specific and highly controlled investigation of the "lure" effects of n-grams from uninverted structures. That is, the experiment asks: "Even though the correct forms *Who can he draw? and Who can he name* are **equally probable statistically,** are uninversion errors more common for the first than the second, since the "lure" bigram *can draw* (**Who he **can draw**?*) is more frequent than the "lure" bigram *can name* (**Who he can name?*)?

With regard to the frequency of the n-grams from uninverted structures, the "high" and "low" frequency (defined with regard to Uninverted Bigram 3) questions in each pair were again perfectly matched for Bigram 1 and Bigram 2 (since the first three words were identical), but **mismatched** as far as possible for Bigram 3 (and likewise for the corresponding trigrams), such that the high-frequency bigram was, in each case, at least 10 times as frequent as the low-frequency bigram. For example,

[2] In these types of circumstances, researchers often report a significance test to show that the "matched" items did not "differ significantly" on the value in question (here, frequency). However, this is not appropriate since such tests are properly used to generalize instances made from a sample to a wider population, and cannot meaningfully be used to draw conclusions about an entire population; here, of test items (Sassenhagen & Alday, 2016).

Table 1. ***Stimulus pairs and n-gram frequencies.***

Target	**Cond.**	**Inverted Trigram1** *who can*	**Inverted Trigram2** *can he draw*	**Inverted Bigram1** *who can*	**Inverted Bigram2** *can he*	**Inverted Bigram3** *he draw*
Who can he draw?	High	6	0	258	850	15
Who can he name?	Low	6	2	258	850	12
What can he eat?	High	42	10	1686	850	323
What can he need?	Low	42	0	1686	850	243
What can he hear?	High	42	13	1686	850	34
What can he mean?	Low	42	2	1686	850	19
Where is Daddy sitting?	High	209	0	34260	578	2
Where is Daddy singing?	Low	209	0	34260	578	2
What can it hold?	High	14	1	1686	226	56
What can it cause?	Low	14	0	1686	226	59
What could it see?	High	24	0	257	158	65
What could it want?	Low	24	0	257	158	50
Why is Daddy hiding?	High	9	0	2469	578	0
Why is Daddy building?	Low	9	0	2469	578	0
What is it wearing?	High	3012	0	87230	19083	0
What is it kissing?	Low	3012	0	87230	19083	0

Target	**Cond.**	**Uninverted Trigram1** *who he can*	**Uninverted Trigram2** *he can*	**Uninverted Bigram1** *who he*	**Uninverted Bigram2** *he can*	**Uninverted Bigram3** *can draw*
Who can he draw?	High	0	6	117	3260	316
Who can he name?	Low	0	1	117	3260	16
What can he eat?	High	33	75	2924	3260	817
What can he need?	Low	33	0	2924	3260	2
What can he hear?	High	33	60	2924	3260	1060
What can he mean?	Low	33	0	2924	3260	9
Where is Daddy sitting?	High	3	1	90	198	579
Where is Daddy singing?	Low	3	0	90	198	57
What can it hold?	High	11	1	2499	954	313
What can it cause?	Low	11	0	2499	954	8
What could it see?	High	7	1	2499	719	313
What could it want?	Low	7	0	2499	719	1
Why is Daddy hiding?	High	1	0	14	198	335
Why is Daddy building?	Low	1	0	14	198	31
What is it wearing?	High	359	0	2499	8081	352
What is it kissing?	Low	359	0	2499	8081	27

Target	**Cond.**	**Unigram1** ***who***	**Unigram2** ***can***	**Unigram3** ***he***	**Unigram4** ***draw***	**Dog's answer**
Who can he draw?	High	41853	102758	212458	5466	His mum!
Who can he name?	Low	41853	102758	212458	6296	His new puppy!
What can he eat?	High	269958	102758	212458	22551	His breakfast!
What can he need?	Low	269958	102758	212458	23302	A new pair of shoes!
What can he hear?	High	269958	102758	212458	7725	A Bird!
What can he mean?	Low	269958	102758	212458	8214	That he is hungry!
Where is Daddy sitting?	High	76055	348124	14295	4212	In the kitchen!
Where is Daddy singing?	Low	76055	348124	14295	3587	In the garden!
What can it hold?	High	269958	102758	260253	8677	A toy
What can it cause?	Low	269958	102758	260253	18436	An accident
What could it see?	High	269958	18299	260253	66313	A mouse!
What could it want?	Low	269958	18299	260253	94362	Cat food!
Why is Daddy hiding?	High	29443	348124	14295	2073	He's playing hide and seek
Why is Daddy building?	Low	29443	348124	14295	1568	He's playing with LEGO
What is it wearing?	High	269958	348124	260253	1550	A sweater!
What is it kissing?	Low	269958	348124	260253	758	Its mum!

considering again the question pair *Who can he draw? / Who can he name?*, with regard to the frequency of the n-grams from uninverted structures, the two are identical with respect to Bigram 1 (*who he*) and Bigram 2 (*he can*), while Bigram 3 is approximately 20 times more frequent for the high-frequency version (*can draw* = 316) than the low-frequency version (*can name* = 12).

In response to presentations of this and previous work, colleagues have often expressed surprise that children hear "uninverted" bigrams (e.g., *who he, he can, can draw*) in the input at all, given that parents and other adults produce few, if any, uninversion errors. It is therefore important to remind the reader that children heard these "uninverted" (with respect to questions) bigrams as part of declarative sentences, including those used for reported speech (e.g., *I wonder* ***who he*** *means;* ***He can*** *do it; You* ***can draw*** *it*). The hypothesis under investigation (which enjoys preliminary support from the study of McCauley et al., 2021) is that, despite having been heard solely in declaratives, these n-grams constitute "lures" towards uninversion errors in question production.

Procedure

The experimenter began the (single) session with the following general instructions:

> Hi, my name is [xxx] and we're going to play a special game with this talking dog. It's a girl dog, and her name is Fifi *[note: this was to ensure that "he" when used in the target questions could not refer to the dog]*. We've got some jigsaws here *[Show Warm-up 1a]* but, uh oh, the jigsaws are all missing some pieces so we can't see what's happening. Luckily, Fifi has got the missing pieces, so we can

> ask her what's happening. Then she'll put in the missing pieces. Don't worry, I'm going to help you by telling you what to ask Fifi.

Showing the first warm-up picture onscreen (see Figure 2a; presented via an Open Sesame script; https://osdoc.cogsci.nl), the experimenter continued:

> So, in this first one, we've got a girl called Sarah. Do you know any girls called Sarah? OK, anyway, so here's Sarah. In this jigsaw, she's carrying something. I wonder what Sarah is carrying. Let's ask the dog what Sarah is carrying. Copy me. Say "What is Sarah carrying?" [Note: in the first two warm-up trials, the experimenter invited the child to copy her question verbatim].

After the child's response (*What is Sarah carrying?*), the experimenter activated the "talking dog" toy to have it produce a pre-recorded answer (here, *a book*). At the same time, the missing pieces of the jigsaw appeared onscreen (see Figure 2b).

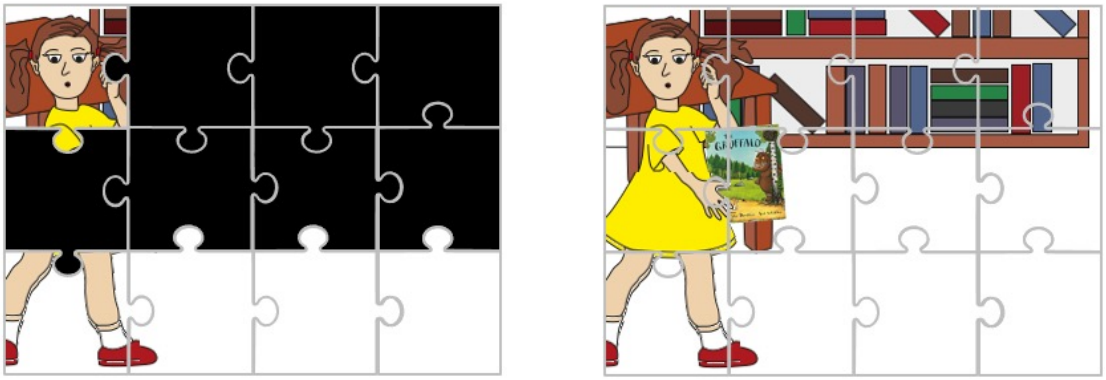

Figure 2. ***Before (2a) and After (2b) pictures shown to children for the first warm-up trial: Q: What is Sarah carrying? A: A book!***

A second warm-up trial (*What is Sarah giving?*) proceeded in the same way, with the child copying the experimenter's question verbatim. For these first two warm-up trials only, the experimenter corrected children who did not produce the target question. For the third warm-up trial, the experimenter announced:

> Now, this time, you're not going to copy me. Instead, I'll just tell you what to ask and you ask it OK? Don't worry, I'll still tell you what to ask. So here's Sarah again. In this jigsaw, she's throwing something. I wonder what Sarah is throwing. Let's ask the dog what Sarah is throwing. You ask the dog what Sarah is throwing.

Note that, for this warm-up trial, and the final, fourth, warm-up trial, the experimenter used indirect/reported speech to present the target question string (grammatically) in uninverted order (*what Sarah is throwing; what Sarah is pushing*). Although the experimenter was careful to always use declarative intonation (i.e., not question intonation), it is important to acknowledge that this method to some extent primes children to produce uninverted questions, both at the abstract level (e.g., *[wh-word] [SUBJECT] [BE] [VERB*]?) and the lexical level (e.g., Savage, Lieven, Theakston, &

Tomasello, 2003; Huttenlocher, Vasilyeva, & Shimpi, 2004; Bencini & Valian, 2008; Rowland, Chang, Ambridge, Pine, & Lieven, 2012). Whether or not this constitutes a confound that potentially invalidates any pattern of uninverted forms found in the data is a question to which we return in the Discussion.

Thus (amongst other possible responses) children could repeat the sequence produced by the experimenter verbatim, yielding an uninversion error, or "invert" the experimenter's question, yielding a correct response. From the third warm-up trial onwards, the experimenter did not correct children's questions, providing only general encouragement. After the final warm-up trial, the experimenter said "Brilliant! OK, now let's try some more pictures with different people in", and proceeded to the 16 test trials, which worked in the same way as the final two warm-up trials. The prompts for the test trials can be found in Appendix 3. Note that while, for warm-up trials, the SUBJECT was always *Sarah*, for the test trials, the SUBJECT was always *he*, *Daddy* or *it*.

In order to sufficiently separate the presentation of the high- and low-frequency (with regard to Uninverted Bigram 3) members of each question pair, the 16 trials were divided into two blocks of 8, presented consecutively. For each participant, two pseudo-randomized lists were created such that if the high-frequency member of a particular question pair appeared in Block 1 (N=4) the low-frequency member of that pair appeared in Block 2 (N=4), and vice versa for the remaining 4 pairs. Within each block, the order of presentation was fully randomized.

Results

We first present the results of our main pre-registered analysis before presenting a number of exploratory analyses designed to investigate the role of the frequency of particular n-grams.

Main, pre-registered analysis

The pre-registered analysis was designed to test the prediction that participants will produce more uninversion errors when those errors incorporate – in the Bigram 3 position – high-frequency bigrams from uninverted structures (e.g., **Who he **can draw**?*) than lower-frequency bigrams from uninverted structures (e.g., **Who he **can name?***). Importantly, when testing this prediction, all other bigrams and unigrams (i.e., single words) are either identical (e.g., *Who+he, he+can, he, can*) or closely matched for frequency (e.g., *draw* and *name* [as verbs] are of approximately equal corpus frequency). Note that, for this analysis, the frequency of the crucial bigram is treated as a categorical predictor (Condition: high/low) since it is manipulated within each otherwise-closely-matched target question pair. Exploratory analyses presented below investigate continuous frequency effects.

Thus, the following pre-registered mixed-effects models syntax[3], for the lme4 package (Bates, Mächler, Bolker & Walker, 2015) in the R environment (R Core Team,

[3] In fact, this model is not quite optimal given the study design, as it fails to take account the fact that TargetSentence is nested inside sentence pair (the pair of sentences matched for n-gram frequency other than the target one). In fact, a model including random slopes for both TargetSentence and

2022), was designed to test the hypothesis of a main effect of condition (high/low frequency, as above), while controlling for children's age in months (scaled and centered) and the potential interaction between these two factors:

glmer(Response ~ Condition * Age + (1+Condition|Subject) + (1+Age|TargetSentence), family="binomial", data=Data)

Responses were coded as (1) uninversion error (e.g., **Who he can draw?*; *N*=159) or (0) correct question (e.g., *Who can he draw?*; *N*=647), with all other responses excluded as missing data (*N*=266); hence the use of a binomial outcome variable (logit function). Although the rate of missing data might seem relatively high, it reflects the fact that – due to our focus on particular n-grams – it was necessary to exclude otherwise-scorable questions where children made perfectly reasonable substitutions (e.g., *Who can the man draw?*). Similar numbers of scorable responses were produced in the high-frequency (*N*=415) and low frequency conditions (*N*=391).

The model set out above failed to converge. Thus, in accordance with our pre-registered analysis plan, we removed the by-TargetSentence random slope of Age, which allowed the model to converge. This model is summarized in Table 2 (see Appendix 1 for the full model). A main effect of Age was observed, reflecting the fact that the rate of uninversion errors decreased with development. However, our pre-registered prediction of a main effect of condition (at $p<0.05$) was not supported; neither was a significant interaction of Condition by Age observed[4]. Indeed, children produced uninversion errors at very similar rates in the high-frequency condition (*M*=0.21, *CI*=0.17-0.25) and the low-frequency condition (*M*=0.19, *CI*=0.15-0.22). Note that the study was powered for a small effect size (*d*=0.2), and so we have reason to consider that this is a genuine null effect rather than a false negative.

Table 2. ***Mixed-effects model for the main, pre-registered analysis. Model summary statistics: AIC=584.6, BIC=622.2, logLik=-284.3, deviance=568.6, df.resid=798.***

Fixed Effect	Estimate	*SE*	*z* value	Pr(>|*z*|)
(Intercept)	-2.98	0.62	-4.78	1.72E-06
ConditionLow	0.24	0.65	0.37	0.7083
Age	-0.85	0.42	-2.01	0.0441
Condition-Low:Age	0.31	0.28	1.12	0.2622

TargetSentencePair failed to converge, apparently because the two are so highly correlated. A model that included a random slope for TargetSentencePair but not TargetSentence yielded similar *p* values to the model reported above, for both Condition ($p=0.44$) and Age ($p=0.04$).

[4] The study pre-registration stated that "P-values will be computed via Kenward-Roger and Satterthwaite approximations". However, this method is in fact applicable for continuous dependent variables only. Thus, we instead report *p* values approximated from the *Z* distribution. We also ran a version of the model with no interaction, in order to allow us to compute *p* values for the main effects of Condition and Age via likelihood ratio test (drop1 function of lme4): $p=0.88$ and $p=0.09$ respectively.

Random effects:

Groups	Name	Variance	*SD*	Corr
Subject	(Intercept)	7.9127	2.813	
	ConditionLow	0.2661	0.5158	-1
TargetSentence	(Intercept)	0.8352	0.9139	

It is important to acknowledge at this point that while this null finding was not predicted by the exemplar-focussed variety of input-based account that we set out to test (e.g., Langacker, 1988; Abbot-Smith & Tomasello, 2006; Goldberg, 2006; Ambridge 2020a, 2020b; McCauley et al., 2021) it is potentially consistent with slot-and-frame-focussed input-based accounts which would seem to assume "free slots" in the crucial Bigram 3 position (e.g., *What+can **[THING] [PROCESS]**?*) (e.g., Rowland & Pine, 2000; Rowland, Pine, Lieven & Theakston, 2003, 2005; Dabrowka & Lieven, 2005; Ambridge, Rowland, Theakston, & Tomasello, 2006; Rowland, 2007; Ambridge & Rowland, 2009). On the other hand, this strict reading of slot-and-frame accounts is difficult to reconcile with the findings of McCauley et al. (2021) of frequency effects in the second, third and fourth bigram positions (spanning the "free slot").

The source of this discrepancy is not easy to pinpoint, but one possibility is that the present dataset does, in principle, include evidence for frequency effects spanning the "free slot', just not necessarily in the third bigram position. We explore this possibility in a series of non-preregistered, exploratory analyses.

Exploratory analyses

The main, pre-registered analysis reported above failed to find any evidence of an effect of the frequency of the third bigram from uninverted structures (e.g., **Who he **can draw/name**?*) on rates of uninversion error. In this analysis, as preregistered, the frequency of the third bigram was treated as a categorical predictor and other n-grams kept constant across the paired items as much as possible. However, there is variance between items beyond these pairs and given that several previous studies have found frequency effects in multiple positions (for n-grams from both inverted and uninverted structures), we conducted a series of exploratory analyses designed to investigate whether any of these effects are observed in the present dataset. Although researcher degrees of freedom are always a concern in non-preregistered analyses (Simmons, Nelson, & Simonsohn, 2011), these are minimized by the fact that our analysis strategy is identical to that of McCauley et al., (2021), with all analyses conducted on the main dataset from the preregistered analysis, with no further exclusions, transformations, recodings etc.

There are various challenges in these analyses, given that the stimuli were not designed to look at these effects, but rather effects within high-/low-frequency matched pairs. Many n-gram frequencies were correlated with one another, creating a problem of multicollinearity. Furthermore, since the present stimuli include just 16 questions (and just 8 *wh*-word+subject+auxiliary combinations), we have a very low ratio of items to predictor variables, which also makes it more difficult to statistically tease apart these predictors (cf., McCauley et al., 2021). Thus, these analyses should be

treated as highly exploratory, and will require confirmation from future suitably designed studies.

In order to address these difficulties, we first took the decision to disregard trigrams, and investigate only the question of whether bigram effects are observed above and beyond unigram (single-word frequency) effects. Excluding trigrams reduces both the problem of collinearity (since trigram frequency is correlated with the frequency of its component unigrams and bigrams) and the low item:predictor ratio (by removing predictors).

We first fit a full model with all unigrams and bigrams (inverted and uninverted) as fixed effects, random effects of participant on the intercept and all slopes, and a random effect of sentence on the intercept. This model did not converge and so we simplified by removing the correlation between the participant random effects. We also excluded uninverted bigram 2 as lme4 determined it to be causing rank-deficiency, presumably because of multicollinearity. This model converged although many of the random effects were returned as zero due to their very small size. To give greater stability throughout our inference process we removed all random effects for slopes that were returned as zero. The random effect of sentence on the intercept was also returned as zero but we retained it in the model in order to be maximally conservative in testing for effects.

In order to see whether any of the n-grams had unique explanatory value with regards to the children's errors, we performed a drop-one analysis where we took the all-predictor model and dropped each n-gram fixed effect in turn, looking at whether doing so hurt fit using a likelihood ratio test. If so then we concluded that it was accounting for unique variance in the full model. The final model is shown in Table 3, which also shows the *p* values from the likelihood ratio (drop1) test. The fixed effects (log_) B1, B2 and B3 refer respectively to the (log-transformed) frequency of the first, second and third bigrams from inverted questions; B1.U, and B3.U of the first and third bigrams from uninverted questions (recall that the second was already excluded earlier). Fixed effects of the (log) frequency of individual words (i.e., unigrams U1, U2, U3 and U4) were included in order to allow us to test whether the frequency of a given individual bigram *combination* explained variance above and beyond the frequency of the individual words that make up that bigram.

This analysis tells us (using the example target question *What are you doing?*) that unigrams 1 and 2 (e.g., *what, are*), inverted bigrams 2 and 3 (*are+you, you+doing*) and uninverted bigram 1 (*what+you*) explain unique variance, with the likelihood of a non-inversion error (the dependent measure) decreasing as a function of the inverted bigram frequency (log_B2, log_B3) and increasing as a function of uninverted bigram frequency (log_B1.U).

Checking for a unique effect of the n-grams is an appropriately conservative way of proceeding. However, is it important to note that, due to collinearity, the absence of a unique effect for any given n-gram could simply be the result of its not being separable from other variables in this particular dataset. In order to look at the theoretical separability of the predictors, we performed Principal Components Analysis (PCA). PCA is a dimensionality-reduction algorithm that, when given a matrix of variables –

Table 3. ***Bigram predictors in exploratory analysis all n-gram model. P values are based on the chi-square (likelihood ratio test) drop-one method. (log_) B1, B2 and B3 refer respectively to the (log-transformed) frequency of the first, second and third bigrams from inverted questions; B1.U, B2.U and B3.U of the first, second and third bigrams from uninverted questions; U1, U2, U3 and U4 of unigrams (i.e., single words). Model summary statistics: AIC=522.9, BIC=593.3, logLik=-246.5, deviance=492.9, df.resid=791.***

Fixed Effect	**M**	**SE**	***p*_drop1**
(Intercept)	-4.1291	0.7314	NA
log_U1	**12.0778**	**3.3484**	**0.0001246**
log_U2	**1.9601**	**1.0172**	**0.04982**
log_U3	-1.7205	1.1222	0.1222
log_U4	1.3875	1.0038	0.1617
log_B1	-0.8387	0.7302	0.2494
log_B2	**-1.883**	**0.5308**	**0.0001334**
log_B3	**-1.927**	**0.8685**	**0.0229**
log_B1.U	**15.2218**	**4.4347**	**0.0002763**
log_B3.U	0.1845	0.1761	0.2925

Random effects:

Groups	**Name**	**Variance**	**Std.Dev.**
Subject	(Intercept)	1.20E+01	3.47E+00
Subject.1	log_U1	4.32E-14	2.08E-07
Subject.2	log_U3	3.27E+00	1.81E+00
Subject.3	log_B1.U	3.39E-01	5.82E-01
TargetSentence	(Intercept)	2.88E-15	5.37E-08

in this case, n-gram predictor variables – collapses highly correlated variables into composite variables ("components"). By looking at how the original variables load onto these components, we can observe how separable they are. Figure 3 shows the loading of all variables onto the first two components, which account for 47% and 27% of the variance respectively. B1, B2 and B3 refer to the first, second and third bigrams from inverted questions; B1.U, B2.U and B3.U to the first, second and third bigrams from uninverted questions; U1, U2, U3 and U4 to unigrams. The further away any two variables are, particularly along the horizontal axis (the first compo accounts for more of the shared variance across the predictors), the more separable they are. It is clear that B1 and B2, being very close together, are hard to separate. It is plausible then that B1 does explain variance in the children's production, but this was a subset of the variance explained by B2, and thus we saw no unique effect of B1. The same applies for B1.U and B2.U, which could explain why B2.U was rejected as rank deficient. A similar situation can be seen for U1 and U3, which could explain why U3 was not found to explain unique variance, and U4 and B3, which could explain why only

the latter explained unique variance. Finally, the very close proximity between U2 and B3.U could explain why only the former is seen to explain unique variance.

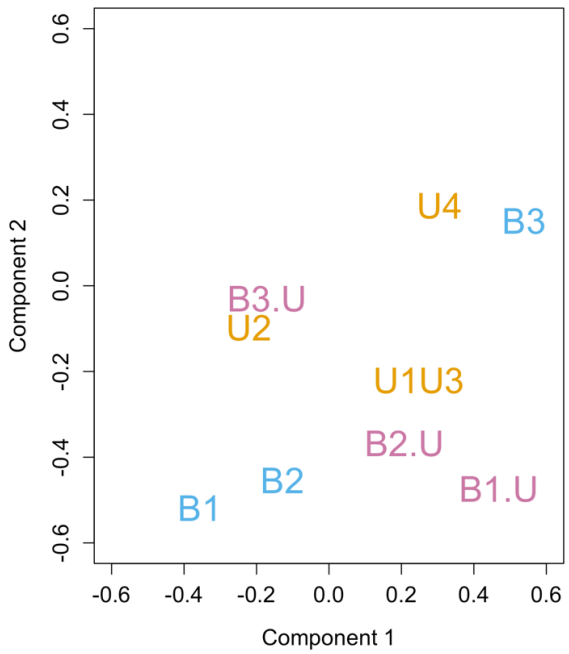

Figure 3: ***Loading of N-gram frequency variables on the first two principal components, which account for 47% and 27% of variance respectively. Unigrams appear in orange (U1, U2, U3, U4), inverted bigrams in light blue (B1, B2, B3) and univerted bigrams in pink (B1.U, B2.U, B3.U). B1, B2 and B3 refer respectively to the first, second and third bigrams from inverted questions; B1.U, B2.U and B3.U to the first, second and third bigrams from uninverted questions; U1, U2, U3 and U4 to unigrams (i.e., single words). The further away any two variables, particularly along the horizontal axis, the greater the extent to which they are separable.***

Summary of Exploratory Effects.

Consistent with a frequent claim in the literature (e.g., Rowland & Pine, 2000; Rowland, 2007; Ambridge & Rowland, 2009), the present exploratory analysis found preliminary evidence that children make fewer uninversion errors for questions that contain bigrams with high input frequency. Although preliminary, this evidence is important, as it is the first experimental study (cf. the corpus study of McCauley et al., 2021) to demonstrate the existence of bigram effects in question production *above-and-beyond effects of the component unigrams*. That is, children make fewer uninversion errors when the bigrams that make up the question (e.g., *can he…*; *…he draw…*) are of higher frequency, independent of the frequency of the individual words (*can, he, draw*). Echoing the corpus study of McCauley et al. (2021), we also found evidence that rates of uninversion error (e.g., **Who he can draw?*) are higher when the uninverted bigrams (e.g., *Who he…*) are of higher frequency in the input. It is important to treat the ***specific*** effects seen with some caution and note that while we saw unique effects of some predictors and not others, the absence of an effect could in part be the effect of collinearity — a variable can spuriously appear not to have an effect

because its variance is being explained by another variable with which it is collinear — and it could be that in a set of stimuli where the variables were more separable we would see different specific patterns. For this reason, we have avoided the temptation of speculating as to potential reasons why it might be *these* particular bigrams that seem to yield frequency effects and not others. Importantly, however, the conclusion that *some* n-gram frequencies are predictive of errors rates is not affected by collinearity, which affects only our ability to say *which ones*.

Discussion

Consistent with input-based accounts of question acquisition, several previous studies have shown that children are less likely to produce uninversion errors (e.g., **Who he can draw?*) when lexical strings that appear in the correct form – particular *wh*-word+auxiliary combinations such as *who can* – are frequent in the input (e.g., Rowland & Pine, 2000; Rowland, Pine, Lieven, & Theakston, 2003, 2005; Dabrowka & Lieven, 2005; Ambridge, Rowland, Theakston, & Tomasello, 2006; Rowland, 2007; Ambridge & Rowland, 2009). Ambridge and Rowland (2009) and McCauley, Bannard, Theakston, Davis, Cameron-Faulkner, and Ambridge (2021) also showed that children are more likely to produce uninversion errors (e.g, **Who he can draw*) when lexical strings that appear in the errorful form are frequent in the input (e.g., *he can* is considerably more frequent than *can he*).

The aim of the present study was to conduct a preregistered experimental test of a prediction that follows from the study of McCauley et al. (2021), and from the more general claim of input-based approaches that learners retain, and are influenced by, individual lexical strings even when they have formed more abstract representations too: Participants will produce more uninversion errors when those errors incorporate – in the Bigram 3 position – high-frequency bigrams from uninverted structures (e.g., **Who he* ***can draw****?*) than lower-frequency bigrams from uninverted structures (e.g., **Who he* ***can name?***); with all other bigrams and unigrams (i.e., single words) either identical (e.g., *Who+he, he+can, he, can*) or closely matched for frequency (e.g., *draw* and *name*). The present study tested this prediction using an elicited-production paradigm in which children put questions to a talking dog toy.

This main, preregistered prediction was not supported. Given that the study was well powered (67 participants, yielding 90% power *a priori*) for even a small effect size (d=0.2), these findings are plausibly consistent with a genuine null effect rather than a false negative. Given that this effect has been seen in naturalistic data (McCauley et al., 2021), it is somewhat surprising that we failed to find it in this experiment. One possible explanation for this discrepancy is that in order to control for the frequency of other component n-grams, we were forced to select items in which there was inadequate difference between the frequency of bigram 3 in the inverted and the uninverted form. On this view, McCauley et al.'s (2021) finding of a frequency effect in (amongst others) the third bigram position reflects a genuine effect, and the present null finding is a result of methodological factors. Of course, it is also possible that the opposite is true: Whenever an effect is found in observational data but not replicated in an experiment, the possibility exists that the apparent effect in the former is due to unmeasured confounding. A third possibility, and the one we favour, is that whether or not frequency effects are observed for a given n-gram position depends

on factors such as the particular lexical question forms under investigation, and participant-level factors such as linguistic history, memory and willingness to generalize beyond the input. In all likelihood, the only way to resolve this issue will be to build detailed computational models that make specific predictions regarding specific lexical question types (possibly for specific individuals), rather than naïve n-gram models that predict equivalent frequency effects across the board.

We follow our pre-registered analysis with non-preregistered exploratory analyses of the data. In this we explored frequency effects for other n-grams. It is important to note that the stimuli were not designed to look at these effects and thus they are confounded with covariance in other n-grams. Nevertheless, we found evidence of a facilitatory effect on correct-question production of the frequency of the second and third bigrams from inverted structures (e.g., *can he…he draw*), even after controlling for unigram frequency (unlike, for example, Ambridge & Rowland, 2009). The frequencies of the first and second bigrams were highly correlated so that it is possible that an effect of the first bigram was hidden. We also saw evidence that rates of uninversion error (e.g., **Who he can draw?*) are higher when the first uninverted bigram (e.g., *Who he…*) are of higher frequency in the input.

Before moving on to explore the potential theoretical implications of the present findings, it is important to acknowledge three possible methodological objections. The first is that – as a result of the tight constraints imposed by the need to match stimuli in the high- and low-frequency conditions – some of the target questions were rather unnatural and/or difficult to illustrate with pictures. It is true that some of the questions are somewhat unnatural, although we did our best to ameliorate any unnaturalness as far as possible with the preliminary lead-in sentences (e.g., *In this jigsaw, it looks like he means something. I wonder what he can mean*). Interested readers are invited to draw their own conclusions regarding the extent to which we succeeded by perusing our full list of prompts, which can be found in Appendix 3 (pictures can be found on the accompanying OSF site at https://osf.io/74urw/. We do not consider it appropriate, however, to conduct an item analysis since our target questions vary with regard to properties other than their perceived naturalness – most importantly the n-gram statistics used as fixed-effect predictors in our exploratory analyses – and one advantage of mixed-effects models is that they allow us to *control for* item-by-item differences that are not captured by the fixed-effects (including naturalness, the particular subject used in the question, differences relating to the illustrations etc.).

The second potential methodological objection is that (as already mentioned in the Methods section), by including uninverted question strings in the experimenter's prompt ("Let's ask the dog *where Daddy is sitting*. You ask the dog *where Daddy is sitting*") we primed children to produce uninverted questions (e.g., **Where Daddy is sitting?*). It is almost certainly the case that such priming will have occurred, since both abstract and lexical priming effects are well established for young children (e.g., Savage et al., 2003; Huttenlocher et al., 2004; Bencini & Valian, 2008; Rowland et al., 2012). The question is whether this priming effect replaced and supplanted children's normal mechanisms of question production to the extent that the present (tentative) findings of certain n-gram effects are entirely invalidated. We do not believe this to be the case for three reasons. First, overall, children produced around four times as many correct as uninverted questions. Clearly, then, children's normal production

mechanisms were, on the whole, operating well; indeed, four times out five, they were able to override any priming effect. Second, although this rate of uninversion errors (20%) is much higher than rates observed in naturalistic data (e.g., McCauley et al., 2021, found just 2%), this is not a fair comparison, since naturalistically-produced questions follow a broadly Zipfian distribution with just a handful of potentially-rote questions (e.g., *What's that?; What are you doing?*) accounting for the majority of all tokens. When we control for this skewed distribution by counting types not tokens, uninversion errors also occur at a rate of around 20% in naturalistic data (e.g., Rowland & Pine, 2000), suggesting that the present method does not artificially inflate rates of universion error; or at least, not to a great extent. Third, at a broad-brush level, the findings of certain n-gram effects on rates of uninversion error echo those of McCauley et al. (2021), which were based entirely on corpus data. Overall, then, we feel justified in claiming that while the experimenter's prompt certainly encouraged children to produce uninversion errors – to some extent, that was exactly the aim – it did so in a way that elucidates, rather than obscures, underlying question-by-question differences in rates of uninversion error versus correct questions.

The third potential methodological objection that we must consider is that by excluding questions that did not use the precise target words (e.g., if the child said, "Who can the man draw?" rather than "Who can he draw?"), we incorrectly estimated overall rates of uninversion errors versus correct questions This is true, but it was never our intention to make any theoretical claims on the basis of *overall* rates of uninversion errors versus correct questions, and, indeed, we do not do so. Any such claim would be problematic given the finding from both the present study (tentatively) and previous studies (more securely) that error rates vary dramatically by question type (e.g., Rowland & Pine, 2000, report uninversion rates of 100% for some questions and 0% for others). Thus, the overall rate of uninversion errors versus correct questions in any particular experimental study is determined, at least to a considerable degree, by the particular question types chosen, meaning that any theoretical claims based on *overall* error rates would invariably be misleading. Relatedly, we do not see it as a problem that particular *wh*-words and particular auxiliaries appeared at unequal rates in our stimuli (which was necessary in order to create closely-matched high-/low-frequency pairs); since at no point do we analyse – much less make claims on the basis of – error rates at the by-*wh*-word or *by*-auxiliary level.

Returning now to the present findings and their implications, when taken together with the findings of McCauley et al. (2021), the exploratory findings from the present study suggest that children's language production mechanism is sensitive to unigram, bigram and trigram frequency, even when those strings are from very different sentence types to the target. That is, strings from declarative input utterances affect the production of questions; specifically, by increasing rates of uninversion error. What is less clear is whether material at the left-hand edge of questions is somehow privileged (e.g., *What are you…*) or, conversely, whether n-grams further to the right from both inverted (e.g., *you doing; doing there*) and uninverted structures (e.g., *you are, are doing, doing there*) play a large – or even larger – role.

Certainly, neither the present findings nor those of McCauley et al. (2021) are consistent with a strict interpretation of proposals such as Rowland and Pine (2000), Dabrowska and Lieven (2005), Ambridge et al. (2006) and Ambridge & Rowland (2009)

that children's early question schemas are of the form *What are [THING] [PROCESS]?* That is, the findings of the present study and McCauley et al. (2021) are not consistent with a "left-edge bias" view under which only the *wh*-word+auxiliary combination is frozen as a learned schema, with the [THING] and [PROCESS] slots entirely "free" (a *strict* interpretation of these previous proposals; and not necessarily an interpretation that their authors would endorse).

In fact, some of the previous evidence for a special role for *wh*-word+auxiliary combinations may not be as strong as it first appears. For example, while Ambridge & Rowland's (2009) experimental study found a negative correlation between children's rates of uninversion error and the input frequency of *wh*-word+auxiliary combinations (or, for *yes/no* questions auxiliary+subject combinations) this correlation held only when removing the outlier Why+can which shows much lower rates of error than would be predicted by its very low input frequency. The corpus study of Rowland and Pine (2000) did not in fact test for this correlation at all, but instead provided evidence only for the weaker claim that "the wh + aux combinations the child uses are more frequent in the mother's input than those the child fails to use (i.e. that occur divided by a subject in uninverted *wh*-questions)". Westergaard (2009) further argues that (1) Many of the child's uninverted questions should have been excluded from Rowland and Pine's (2000) analysis because they were produced only once or a handful of times and (2) Many of the child's inverted questions should have been excluded, because they include the dummy auxiliary DO (e.g., *What does; Where did*) which children already know – for quite independent reasons – is not normally included after a subject unless for emphasis. For example, a child would not normally say *He does like it* or *He did go to school* (cf., *He likes it; He went to school)* rendering the non-occurrence of *What he does like?* or *Where he did go?* moot.

The most convincing evidence for a special role for the left-edge of the utterance comes from the corpus study of Rowland (2007, which found a significant negative correlation between the frequency of the frame (again defined as wh+auxiliary for wh-questions and auxiliary+subject for yes/no questions) and rates of uninversion errors (versus correct questions), over and above auxiliary type (DO vs modal). However, this study did not control for the independent input frequency of other bigrams in the well-formed question, or of unigrams.

Recall, too, that the present study does not constitute strong evidence against a special role for the first bigram (here, always *wh*-word+auxiliary) due to collinearity between the input frequency of the first bigram (e.g., *what+are*), which was not a significant predictor of correct production, and the second bigram (*are+you*), which was. Thus, the jury is still very much out with regard to the question of whether the n-grams at the left edge of the utterance hold some special importance for question acquisition (e.g., by leading to the formation of slot-and-frames patterns like *What are [THING] [PROCESS]*)? It is also important to emphasize that while Rowland and Pine (2000) and Rowland (2007 focussed on the *protective* effect of high-frequency inverted strings on correct question production, the present study (like Ambridge & Rowland, 2009; McCauley et al., 2021) additionally investigated the potentially error-causing (lure) effect of high-frequency uninverted strings on uninversion errors.

What the present exploratory findings tentatively suggest is that at a general level (i.e., setting aside the question of a left-edge bias), n-gram frequency indeed affects the relative probability of uninversion errors versus correct-question production: high-frequency n-grams with inverted order pull towards correct questions; high-frequency n-grams with uninverted word order pull towards non-inversion errors. Thus the present findings – like those of McCauley et al. (2021) – are consistent with a view under which, having generalized in some sense across input utterances to yield more abstract representations, traces left by the initial input utterances are not discarded but retained (in principle, forever), and influence subsequent language production and processing (e.g., Langacker, 1988; Abbot-Smith & Tomasello, 2006; Goldberg, 2006; Ambridge 2020a, 2020b). As discussed in Ambridge (2020b) the generalizations yielded by such a model are likely to exist at numerous levels of abstraction simultaneously, and may look very different to the types of generalizations posited under traditional linguistic analysis (e.g., a *[WH-WORD] [AUXILIARY] [THING] [PROCESS]*? construction or a *subject-auxiliary inversion* rule). Indeed, it is important to emphasize that at a global level (we are not aware of any studies looking specifically at adult question production) frequency effects are ubiquitous not just in child language acquisition (e.g., Ambridge et al., 2015), but in adult language processing too (in addition to the studies cited in the Introduction, see e.g., the summaries by Ellis, 2002; Gries & Divjak, 2012). Frequency effects – including n-gram effects – are not solely a hallmark of child language acquisition that disappear when more abstract representations are formed. Rather, what we need are accounts that can explain both abstract and lexical effects at once, for both adults and children.

On this note, it is important to reiterate, as stated in the Introduction, that the present findings (and McCauley et al., 2021) do not demonstrate the *absence* of a syntactic subject-auxiliary inversion rule. What they do suggest is that proponents of such accounts owe an explanation as to the source of the observed unigram, bigram, and trigram effects; for example, in terms of the filtering of a *subject-auxiliary inversion* rule through a production mechanism that is sensitive to n-gram frequency. Note that this is only a suggestion; we are not aware of any rule-based accounts of the acquisition of question production that actually incorporate such a mechanism.

In turn, researchers who advocate the abandoning of accounts based on the notion of a subject-auxiliary inversion rule owe an account of exactly how children acquire the ability to move beyond the n-gram strings that they hear in the input and develop abstract representations that allow them to produce entirely novel questions (including those for which many individual n-gram frequencies will be zero, or at least extremely low).

At present, descriptive verbal accounts – on both the rule-based and construction-based sides – do not make sufficiently precise quantitative predictions that they can be subjected to objective empirical testing. For example, as we have noted throughout, it is not clear whether slot-and-frame-based accounts really predict the absence of frequency effects in "free slot" position (e.g., *What+can [THING] [PROCESS]?*), or even – necessarily – their attenuation. If precise quantitative predictions are to be derived from accounts of question acquisition, then it will almost certainly be necessary to implement these accounts as mechanistic computational models.

In the meantime, while the present study – contra McCauley et al. (2021) – found no evidence for the special importance in question formation of the third bigram from uninverted utterances, it does suggest that children's question production is indeed influenced by unigram, bigram, and trigram frequency; findings that any successful account of children's question acquisition – and of their language acquisition more generally – will need to explain.

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, *23*(3), 275-290.

Ambridge, B. (2020a). Against stored abstractions: A radical exemplar model of language acquisition. *First Language, 40*(5-6), 509-559.

Ambridge, B. (2020b). Abstractions made of exemplars or 'You're all right and I've changed my mind' Response to commentators. *First Language, 40*(5-6), 640-659.

Ambridge, B., & Rowland, C.F. (2009). Predicting children's errors with negative questions: Testing a schema-combination account. Cognitive Linguistics, 20(2), 225-266.

Ambridge, B., Rowland, C. F., & Pine, J. M. (2008). Is Structure Dependence an innate constraint? New experimental evidence from children's complex question production. Cognitive Science, 32(1), 222-255.

Ambridge, B., Rowland, C. F., Theakston, A. & Tomasello, M. (2006) Comparing Different Accounts of Non-Inversion Errors in Children's Non-Subject Wh-Questions: 'What experimental data can tell us?' *Journal of Child Language* 30(3) 519-557.

Ambridge, B., Rowland, C.F., Theakston, A.L. & Kidd, E.J. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language, 42*(2), 239-73.

Arnon, I. & Clark, E. V. (2011). When 'on your feet' is better than 'feet': Children's word production is facilitated in familiar sentence-frames. *Language Learning and Development, 7*, 107-129.

Arnon, I. & Snider, N. (2010) More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*: 67–82.

Arnon, I., McCauley, S.(C) & Christiansen, M. H. (2017). Digging up the building blocks of language: Age-of-Acquisition effects for multiword phrases. Journal of Memory and Language, 92, 265-280.

Bannard, C. (2006). *Acquiring phrasal lexicons from corpora* (Doctoral dissertation, University of Edinburgh).

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science, 19*(3), 241-248.

Bates, D., Mächler, M., Bolker, B. & Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software, 67*(1), 1–48. doi: 10.18637/jss.v067.i01.

Bellgui, U. (1971). Simplification in children's language. In R. Huxley & E. In-gram (Eds.), *Language acquisition: Models and methods*. New York: Academic Press.

Bencini, G. M. L., & Valian, V. V. (2008). Abstract sentence representations in 3 year-olds: Evidence from language production and comprehension. *Journal of Memory and Language, 59*, 97 - 113.

Bloom, L., Merkin, S., & Wooten, J. (1982). Wh-Questions: Linguistic factors that contribute to the sequence of acquisition. *Child Development, 53*, 1084-1092.

Bloom, L., Merkin, S., & Wooten, J. (1982). Wh-Questions: Linguistic factors that contribute to the sequence of acquisition. *Child Development, 53*, 1084-1092.

Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: the reduction of don't in English. *Linguistics, 37*(4), 575-596.

Dabrowska, E., & Lieven, E. (2005). Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics, 16*(3), 437-474.

DeVilliers, J. (1991). Why question? In T. L. Maxfield & B. Plunkett (Eds.), *Papers in the acquisition of wh: Proceedings of the UMASS Roundtable, May 1990*. Amherst, MA: University of Massachusetts Occasional Papers.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition, 24*(2), 143-188.

Erreich, A. (1984). Learning how to ask: Patterns of inversion in yes-no and wh-questions. *Journal of Child Language, 11*, 597-592.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. New York: Oxford University Press.

Gries, S. T., & Divjak, D. (Eds.). (2012). *Frequency effects in language learning and processing*. De Gruyter Mouton.

Hattori. (2003). Why do children say did you went?: the role of do-support. *Supplement to the Proceedings of the 28th Boston University Conference on Language Development*. (http://www.bu.edu/linguistics/APPLIED/BUCLD/supp.html)

Havron, N., & Arnon, I. (2021). Starting big: The effect of unit size on language learning in children and adults. *Journal of Child Language*, *48*(2), 244-260.

Hernández, M., Costa, A., & Arnon, I. (2016). More than words: multiword frequency effects in non-native speakers. *Language, Cognition and Neuroscience*, *31*, 785-800.

Hurford, J. (1975). A child and the English question formation rule. *Journal of Child Language, 2*, 299-301.

Huttenlocher, J., Vasilyeva, M., & Shimpi, P. (2004). Syntactic priming in young children. *Journal of Memory and Language, 50*(2), 182-195.

Janssen, N. & Barber, H.A. (2012) Phrase frequency effects in language production. PLoS ONE 7(3): e33202. doi:10.1371/journal.pone.0033202.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. in Bybee, Joan and Paul Hopper (eds.). 2000. Frequency and the emergence of linguistic structure. Amsterdam: John Benjamins (pp. 229-254).

Kueser, J.B, & Leonard, L.B. (2020. The Effects of frequency and predictability on repetition in children with Developmental Language Disorder. *Journal of Speech Language and Hearing Research, 63*(4):1165-1180. doi: 10.1044/2019_JSLHR-19-00155.

Krug, M. (1998). String frequency. A cognitive motivating factor in coalescence, language processing, and linguistic change. *Journal of English Linguistics, 26*, 286–320.

Kuczaj, S. (1976). Arguments against Hurford's 'Aux copying rule'. *Journal of Child Language, 3*, 423-427.

Kuczaj, S. A., & Brannick, N. (1979). Children's use of the wh question modal auxiliary placement rule. *Journal of Experimental Child Psychology, 28*, 43-67.

Labov, W., & Labov, T. (1978). Learning the syntax of questions. In R. Campbell & P. Smith (Eds.), *Recent advances in the psychology of language*. New York: Plenum Press.

Langacker, R.W., (1988). A usage-based model. In B. Rudzka-Ostyn (ed.), Topics in Cognitive Linguistics. Amsterdam: John Benjamins, pp. 127-161.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech, 6*(3), 172-187.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database.* Mahwah, NJ: Lawrence Erlbaum Associates.

Maratsos, M., & Kuczaj, S. (1978). Against the transformationalist account: A simpler analysis of auxiliary overmarking. *Journal of Child Language, 5*, 337-345.

McCauley, S. M., Bannard, C., Theakston, A., Davis, M., Cameron-Faulkner, T., & Ambridge, B. (2021). Multiword units lead to errors of commission in children's spontaneous production: "What corpus data can tell us?*" *Developmental Science, 24*(6), e13125.

McDonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities. *Psychological Science, 14*, 648–652.

Pluymaekers, M., Ernestus, M., & Baayen, R. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica, 62*(2-4), 146-159.

Pozzan, L., & Valian, V. (2017). Asking questions in child English: Evidence for early abstract representations. *Language Acquisition, 24*(3), 209-233.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ringstad, T. & Kush, D. (2021) Learning embedded verb placement in Norwegian: Evidence for early overgeneralization. *Language Acquisition.*

Rowland, C. F. (2007). Explaining errors in children's questions. *Cognition*, *104*(1), 106-134.

Rowland, C. F., & Pine, J. M. (2000). Subject-auxiliary inversion errors and wh-question acquisition: 'What children do know?' *Journal of Child Language, 27*(1), 157-181.

Rowland, C. F., Pine, J. M., Lieven, E. V. M., & Theakston, A. L. (2003). Determinants of acquisition order in wh questions: re-evaluating the role of caregiver speech. *Journal of Child Language,* 609-635.

Rowland, C.F., Pine, J. M., Lieven, E. V. M., & Theakston, A. L. (2005). The incidence of error in young children's wh-questions. *Journal of Speech, Language and Hearing Research, 48* 384-404.

Rowland, C.F., Chang, F., Ambridge, B., Pine, J. M., & Lieven, E.V.M. (2012). The development of abstract syntax: Evidence from structural priming and the lexical boost. *Cognition, 125*(1) 49–63.

Santelmann, L., Berk, S., Austin, J., Somashekar, S., & Lust, B. (2002). Continuity and development in the acquisition of inversion in yes/no questions: dissociating movement and inflection. *Journal of Child language, 29*(4), 813-842.

Sassenhagen, J., & Alday, P. M. (2016). A common misapplication of statistical inference: Nuisance control with null-hypothesis significance tests. *Brain and Language, 162,* 42-45.

Savage, C., Lieven, E., Theakston, A., & Tomasello, M. (2003). Testing the abstractness of children's linguistic representations: lexical and structural priming of syntactic constructions in young children. *Developmental Science, 6*(5), 557-567.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science, 22*(11), 1359-1366.

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W.J.B. (2011) Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition 37*: 776.

Skarabela, B., Ota, M., O'Connor, R., & Arnon, I. (2021). 'Clap your hands' or 'take your hands'? One-year-olds distinguish between frequent and infrequent multiword phrases. *Cognition, 211,* 104612.

Sosa, A.V. & MacFarlane, J. (2002) Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word of. *Brain and Language 83*: 227–236.

Stromswold, K. (1990). *Learnability and the acquisition of auxiliaries.* Unpublished Ph.D. dissertation, MIT.

Stromswold, K. (1995). The acquisition of subject and non-subject wh-questions. *Language Acquisition, 4*(1), 5-48.

Tremblay, A. and Baayen, R. H. (2010) Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood, *Perspectives on formulaic language: Acquisition and communication*. London: The Continuum International Publishing Group

Tyack, D., & Ingram, D. (1976). Children's production and comprehension of questions. *Journal of Child Language, 4*, 211-224.

Valian, V., & Casey, L. (2003). Young chilren's acquisition of wh-questions: the role of structured input. *Journal of Child Language, 30,* 117-143.

Valian, V., Lasser, I., & Mandelbaum, D. (1992). *Children's early questions.* Paper presented at the 17th Annual Boston University Conference on Language Development, Boston, MA.

Waldmann, C. (2011). Moving in small steps towards verb second: A case study. *Nordic Journal of Linguistics, 34*(3), 331-359.

Westergaard, M. (2009). Usage-based vs. rule-based learning: the acquisition of word order in wh-questions in English and Norwegian. *Journal of Child Language, 36*(5), 1023-1051.

Westergaard, M. & K. Bentzen. (2007). The (non-) effect of input frequency on the acquisition of word order in Norwegian embedded. *Frequency effects in language acquisition: Defining the limits of frequency as an explanatory concept*, *32*, 271.

Data, code and materials availability statement

All raw data, analysis code (for R) and materials (a package for the Open Source Python package, Open Sesame: https://osdoc.cogsci.nl) can be downloaded from https://osf.io/74urw/

Ethics statement

Ethics approval was obtained from the University of Liverpool Research Ethics Committee prior to recruitment. Children's caregivers gave informed written consent and children gave verbal assent.

Authorship and Contributorship Statement

BA, SM, CB, TC-F and AT conceived of the study and designed the study. SM and BA wrote the first draft of the manuscript. AG contributed to the design of the study and collected the data. SM, CB and BA analyzed the data. BA, SM, CB and AT revised the manuscript. All authors approved the final version of the manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Acknowledgements

This work was supported by the International Centre for Language and Communicative Development (LuCiD). The support of the Economic and Social Research Council [ES/L008955/1 and ES/S007113/1] is gratefully acknowledged.

Appendix 1: Full R output for the main preregistered analysis

```
[1] "Now the preregistered model: We said 'In the event of convergence
    failure, we will simplify the model by simplifying the random effects
    terms to no longer include the by-subject random slope for condition
    or the by-item random slope for age. In the event of further conver-
    gence failure we will remove the fixed effect of subject age'"
[1] "Here's a summary of the final model - We had to drop the by-Tar-
    getSentence random slope for Age"
Generalized linear mixed model fit by maximum likelihood (Laplace Approxi-
    mation) ['glmerMod']
 Family: binomial  ( logit )
Formula: Response ~ Condition * Age + (1 + Condition | Subject) + (1 |
    TargetSentence)
   Data: Data

     AIC      BIC   logLik deviance df.resid
   584.6    622.2   -284.3    568.6      798

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9240 -0.2798 -0.1391 -0.0620  4.9796

Random effects:
 Groups         Name         Variance Std.Dev. Corr
 Subject        (Intercept)  7.9127   2.8130
                ConditionLow 0.2661   0.5158   -1.00
 TargetSentence (Intercept)  0.8352   0.9139
Number of obs: 806, groups:  Subject, 67; TargetSentence, 16

Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.9802     0.6229  -4.784 1.72e-06 ***
ConditionLow       0.2448     0.6543   0.374   0.7083
Age               -0.8488     0.4217  -2.013   0.0441 *
ConditionLow:Age   0.3144     0.2804   1.121   0.2622
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
optimizer (Nelder_Mead) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')

[1] "In the preregistration, we said 'P-values will be computed via Ken-
    ward-Roger and Satterthwaite approximations', but this isn't actually
    possible for binomial models So we'll just report the p values from
    the main model output (approximated via the z distribution"
[1] "As a double-check, we'll remove the interaction, which will allow us
    to get p values via drop1, and report this in a footnote"
Single term deletions

Model:
Response ~ Condition + Age + (1 + Condition | Subject) + (1 |
    TargetSentence)
          npar    AIC     LRT Pr(Chi)
<none>         583.93
Condition    1 581.95 0.02461 0.87534
Age          1 584.82 2.89179 0.08903 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
[1] "Now, the pre-registered syntax is all very well, but it seems to me
    (Ben) that we should also include pair ('Set') as a random effect,
    since the high/low frequency manipulation is indeed within each pair,
    and again report it in a footnote"
[1] "Just fails as they're too correlated"
[1] "Probably makes more sense than the pre-registered syntax, but doesn't
    actually change the result at all"
Generalized linear mixed model fit by maximum likelihood (Laplace Approxi-
    mation) ['glmerMod']
 Family: binomial  ( logit )
Formula: Response ~ Condition * Age + (1 + Condition | Subject) + (1 |
    Set)
   Data: Data

     AIC      BIC   logLik deviance df.resid
   572.3    609.8   -278.2    556.3      798

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9853 -0.2855 -0.1379 -0.0594  5.2268

Random effects:
 Groups  Name         Variance Std.Dev. Corr
 Subject (Intercept)  8.3877   2.8962
         ConditionLow 0.3166   0.5627   -1.00
 Set     (Intercept)  0.9492   0.9743
Number of obs: 806, groups:  Subject, 67; Set, 8

Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.0690     0.6402  -4.794 1.64e-06 ***
ConditionLow       0.3380     0.4344   0.778   0.4365
Age               -0.8952     0.4320  -2.072   0.0382 *
ConditionLow:Age   0.3540     0.2775   1.276   0.2020
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix 2: Full R output for the exploratory analyses

```
[1] "First model includes the frequency of all unigrams - i.e.,  each of the individual words - called
    log_U1/U2/U3/U4 - all bigrams from the inverted form of the question, called log_B1/B2/B2, and all bigrams from
    the uninverted form of the question, called log_B1.U/B2.U/B3.U. We attempt to include a by-participant random
    slope for all of these predictors, but as we'll see later this won't converge. There are no possible by-Tar-
    getSentence random slopes"
[1] "Doesn't converge so simplify - starting by removing the correlations between the random effect of structure.
    Also remove B2.U as glmer rejects: fixed-effect model matrix is rank deficient"
[1] "Now converges but gives a singular fit. To improve stability for model comparisons, remove all random effects
    that explain close to zero variance (shows up as 0.000e+00) except for TargetSentence which we retain for rea-
    sons of conservatism"
[1] "Still a singular fit, but that's OK!"
[1] "# Now perform a drop one analysis to look at unique contribution of each of the n-grams"
Data: Data
Models:
M_U1: Response ~ log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_U1   14 535.64 601.33 -253.82   507.64
M      15 522.92 593.30 -246.46   492.92 14.721  1  0.0001246 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Unigram1 is retained in the final model"
Data: Data
Models:
M_U2: Response ~ log_U1 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_U2   14 524.77 590.46 -248.38   496.77
M      15 522.92 593.30 -246.46   492.92 3.8475  1    0.04982 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Unigram2 is (narrowly!) retained in the final model"
```

```
Data: Data
Models:
M_U3: Response ~ log_U1 + log_U2 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
    npar    AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_U3   14 523.31 589.0 -247.66   495.31
M      15 522.92 593.3 -246.46   492.92 2.3891  1     0.1222
[1] "Uingram3 is NOT retained in the final model"
Data: Data
Models:
M_U4: Response ~ log_U1 + log_U2 + log_U3 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_U4   14 522.88 588.57 -247.44   494.88
M      15 522.92 593.30 -246.46   492.92 1.9582  1     0.1617
[1] "Uingram4 is NOT retained in the final model"
Data: Data
Models:
M_B1: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_B1   14 522.25 587.94 -247.12   494.25
M      15 522.92 593.30 -246.46   492.92 1.3268  1     0.2494
[1] "Bigram1 from INVERTED forms is NOT retained in the final model"
Data: Data
Models:
M_B2: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
    npar    AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_B2   14 535.51 601.2 -253.76   507.51
```

```
M      15 522.92 593.3 -246.46   492.92 14.593  1  0.0001334 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Bigram2 from INVERTED forms IS retained in the final model"
Data: Data
Models:
M_B3: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
     npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_B3   14 526.10 591.79 -249.05   498.10
M      15 522.92 593.30 -246.46   492.92 5.1758  1     0.0229 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Bigram3 from INVERTED forms IS retained in the final model"
Data: Data
Models:
M_B1.U: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B3.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_B1.U   14 534.15 599.84 -253.07   506.15
M        15 522.92 593.30 -246.46   492.92 13.224  1  0.0002763 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Bigram1 from UNinverted forms IS retained in the final model"
Data: Data
Models:
M_B3.U: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + (1 + log_U1 + log_U3 +
    log_B1.U || Subject) + (1 | TargetSentence)
M: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 + log_B3 + log_B1.U + log_B3.U + (1 + log_U1 +
    log_U3 + log_B1.U || Subject) + (1 | TargetSentence)
       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
M_B3.U   14 522.03 587.72 -247.01   494.03
M        15 522.92 593.30 -246.46   492.92 1.1082  1     0.2925
[1] "Bigram3 from UNinverted forms IS NOT retained in the final model"
```

```
[1] "Recall that Bigram2 from UNinverted forms IS NOT retained in the final model as it was already dropped due to
    colinearity"
[1] "Summary: U1, U2, B2, B3 and B1.U explain unique variance"
[1] "Next do a PCA of the bigrams to understand what is going on"
[1] "principal package doesn't do simple PCA. It does PCA plus rotation, so switching to prcomp which is a built-in
    R function"
[1] "Here's the model summary for Table 3"
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: Response ~ log_U1 + log_U2 + log_U3 + log_U4 + log_B1 + log_B2 +
    log_B3 + log_B1.U + log_B3.U + (1 + log_U1 + log_U3 + log_B1.U ||      Subject) + (1 | TargetSentence)
   Data: Data
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
   522.9    593.3   -246.5    492.9      791

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.7235 -0.1977 -0.0735 -0.0191 10.8171

Random effects:
 Groups         Name        Variance  Std.Dev.
 Subject        (Intercept) 1.201e+01 3.466e+00
 Subject.1      log_U1      4.322e-14 2.079e-07
 Subject.2      log_U3      3.265e+00 1.807e+00
 Subject.3      log_B1.U    3.385e-01 5.818e-01
 TargetSentence (Intercept) 2.880e-15 5.367e-08
Number of obs: 806, groups:  Subject, 67; TargetSentence, 16

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.1291     0.7314  -5.645 1.65e-08 ***
log_U1      -12.0778     3.3484  -3.607 0.000310 ***
log_U2        1.9601     1.0172   1.927 0.053991 .
log_U3       -1.7205     1.1222  -1.533 0.125230
log_U4        1.3875     1.0038   1.382 0.166924
log_B1       -0.8387     0.7302  -1.149 0.250691
```

```
log_B2       -1.8830     0.5308  -3.548 0.000389 ***
log_B3       -1.9270     0.8685  -2.219 0.026510 *
log_B1.U     15.2218     4.4347   3.432 0.000598 ***
log_B3.U      0.1845     0.1761   1.047 0.294886
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
         (Intr) log_U1 log_U2 log_U3 log_U4 log_B1 log_B2 log_B3 l_B1.U
log_U1    0.272
log_U2   -0.108 -0.296
log_U3    0.112  0.777 -0.096
log_U4   -0.062 -0.289  0.876 -0.065
log_B1    0.064  0.023 -0.700  0.227 -0.516
log_B2    0.235  0.629 -0.187  0.239 -0.040 -0.231
log_B3    0.133  0.575 -0.781  0.357 -0.812  0.533  0.381
log_B1.U -0.259 -0.988  0.361 -0.820  0.323 -0.152 -0.594 -0.632
log_B3.U -0.045 -0.118  0.408  0.015  0.468 -0.265 -0.094 -0.432  0.142
optimizer (bobyqa) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

Appendix 3: Full text of all prompt questions

Oh look…	**In this jigsaw…**	**I wonder/Let's ask the dog/You ask the dog**
… here's the BOY	… he's naming someone	…who he can name
… here's the BOY	… he's drawing someone	…who he can draw
… here's the BOY	… he always needs something	…what he can need
… here's the BOY	… he always eats something	…what he can eat
… here's the BOY	… it looks like he means something	…what he can mean
… here's the BOY	… it looks like he hears something	…what he can hear
… here's DADDY	… he's singing somewhere	…where Daddy is singing
… here's DADDY	… he's sitting somewhere	…where Daddy is sitting
… here's the CAT	… it's causing something	…what it can cause
… here's the CAT	… it's holding something	…what it can hold
… here's the CAT	… it looks like it wants something	…what it could want
… here's the CAT	… it looks like it sees something	…what it could see
… here's DADDY	… he's building, for some reason	…why Daddy is building
… here's DADDY	… he's hiding, for some reason	…why Daddy is hiding
… here's the CAT	… it's kissing something	…what it is kissing
… here's the CAT	… it's wearing something	…what it is wearing

License