# Adults Behaving Childishly: Errors in Adult Responses to *Wh*-Questions

C. Jane Lutken and Karin Stromswold*

## 1 Introduction

Errors are often used as windows into the language processing mechanism. For example, over-regularization errors in young children have been interpreted as evidence that they know and use the rules of their target language but have misapplied a rule when faced with an exception (e.g. Brown 1973). However, not all errors can be explained as over-regularizations of the target grammar. One error is in answering questions with multiple clauses. When asked a question with a *wh*-relativizer (as in (1)), English-speaking kindergarteners sometimes incorrectly answer the *wh*-relativizer, rather than the matrix *wh*-word (e.g. de Villiers and Roeper 1995, de Villiers, Kotfila, and Klein 2019, Thornton and Crain 1994, Lutken, Legendre and Omaki, 2020). For example, English-speaking children sometimes respond to (1) by answering what Lewis said he picked (apples) rather than how Lewis did the telling (on the phone). In other words, they apparently interpret (1) to mean (2).

(1) How did Lewis tell Sally what he picked?
(2) What did Lewis tell Sally (that) he picked?

These errors are prevalent and replicated in experiments. English-speaking children's misinterpretations are particularly interesting because they are superficially similar to the *wh*-scope marking (WSM) construction found in languages like German where a true *wh*-phrase appears medially while the scope of the *wh*-phrase is marked by an initial, contentless *wh*-phrase, or scope marker (Lutz, Muller, and von Stechow 2000). In the German question (3), for example, *was* serves as a scope marker while *wer* is the contentful *wh*-phrase that the child should answer.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (3) | Was | glaubst | Du, | wer | die | gute | Fee | ist? |
| | What | think | you | who | the | good | fairy | is? |
| | 'Who do you think the good fairy is?' | | | | | | | |

Crucially, German questions like (3) are syntactically distinct from English questions like (1): *wer* is a contentful *wh*-phrase, not a relativizer. Thus, in these German questions, the appropriate answer is a response to the medial *wh*-word. The similarity between English questions like (1) and German questions like (3) is purely superficial: both follow a *wh…wh…* pattern.

Why English-speaking children sometimes treat medial *wh*-relativizers as true *wh*-phrases is theoretically interesting and important. It could be the result of immature syntactic competence, with the English-speaking child temporarily using an alternative grammar that is licensed by Universal Grammar (UG). Under either a Principles and Parameters or an Optimality Theory model, the innate language faculty provides access to potential syntactic variations and learners eventually select the right grammar for their language based on exposure. The innate access to UG allows for temporary adoption of the incorrect parameter setting or constraint ranking while choosing among competing grammars (e.g. Yang 2002, Legendre, Vainikka, Hagstrom and Todorova 2002, Snyder 2007). These errors could be due to English-speaking children having immature competence, in which they erroneously assume English utilizes WSM and, thus, they treat the matrix *wh*-phrase as a scope marker and the medial *wh*-phrase as the one they should answer. De Villiers and Roeper (1995) suggest that this is the cause of these errors.

On the other hand, English-speaking children might have the correct adult-like grammar for English questions but have immature processing mechanisms that cause them to make these errors. Omaki and Lidz (2015), Snedeker (2013) and others have argued that certain types of children's

*We thank the Rutgers University students who participated in our experiments. We also gratefully thank the members of the Stromswold lab for useful feedback and comments, in particular to the research assistants who assisted in the coding of the data in Experiments 1 and 2. This work was supported by a Rutgers Research Faculty Grant to Karin Stromswold.

errors are due to children having more limited processing abilities than adults, rather than children having different grammars than adults. Lutken, Legendre, and Omaki (2020) argue that this is the cause of English-speaking children's misinterpretation of questions like (1): they are not *true* cases of WSM, but merely *resemble* WSM. Consistent with this, Lutken and Legendre (2021) found a relationship between children's working memory (WM) and the rate of their WSM-like interpretations: children with smaller WMs made more WSM-like interpretations.

The current study tests the predictions of the immature competence and immature processing accounts of English-speaking children's errors by investigating the types of errors English-speaking adults make in response to complex questions when their processing abilities are overtaxed. The two accounts make distinct predictions. If English-speaking children's errors are due to them having immature grammars (the immature competence account), English-speaking adults should never make WSM errors because adults' grammars are fully developed. No matter how great the processing load, adults should not revert to a non-English grammar. However, if children's WSM-like behavior is the result of their language processor being overtaxed (the immature processing account), adults might also make WSM-like errors when they are overtaxed. Thus, testing English-speaking adults' understanding of questions with *wh*-relativizers when their processors are overtaxed can provide insight into why English-speaking children make WSM errors.

We conducted 2 online comprehension experiments designed to test whether adults make WSM errors when they are overtaxed. In previous experiments that tested children's understanding of complex questions (e.g. de Villiers and Roeper 1995, Lutken et al. 2020, Lutken and Legendre 2021), children's interpretation of questions was assessed by having them answer questions about short stories. In these experiments, stories and questions were accompanied by cartoon illustrations and experimenters used gestures, repetitions, and pointing when they presented the stories and questions. In addition, experimenters provided clarifications as needed. Unsurprisingly, when given the same tasks as the children, adult controls in these experiments never made WSM errors. The experiments presented in this paper use same question-after-story format as in previous studies, but adults received no clarifications, gestures, repetitions, or visual cues. These adult experiments differed from previous experiments in several other ways. First, the task was written rather than spoken. Second, the stories were longer and more complex (mean length = 134 words; 8.3 sentences). Finally, the questions' matrix and embedded verbs were more abstract.

The results of these experiments reveal that, not only do adults make WSM-like errors when they are given an appropriately difficult task, they do so at a similar rate as children do, thus supporting the hypothesis that these errors in "child" English are the result of immature processing mechanisms rather than immature competence.

## 2 Experiments

### 2.1 Participants

Twenty-four adults participated in a pilot experiment, 47 adults participated in Experiment 1, and 24 adults participated in Experiment 2. All participants were Rutgers College students who reported being native speakers of English. All students received course credit for their participation and were only allowed to participate in only one of the three experiments. Participants completed a demographics and language background questionnaire. We excluded participants who reported having a history of a hearing impairment or language impairment. In both experiments and the pilot, participants were allowed to give multiple responses to questions. However, if a participant consistently selected all of the options, we did not include the participant's data in any analyses. Experiment 1 included a working memory (WM) task, and we excluded participants who got more than half of the WM items wrong. Similarly, Experiments 1 and 2 included 6 "catch" trials with simple content questions, and participants who got more than half of the catch trials wrong were excluded.

### 2.2 Stimuli

Stimuli consisted of a short, context-establishing story followed by a question. Target questions were of the form "*wh… wh…?",* containing a true *wh*-word in the matrix position and a *wh*-relativizer in the medial position. Thus, questions resembled those used in child studies. However, both stories and questions differed from those used in previous work as described below.

### 2.2.1 Question & Story Design

The verbs in the questions were specifically selected to be more abstract than those used in previous work. In the current experiments, the two matrix clause verbs were *report* and *confess*. Like *say* and *tell* (which were the matrix verbs in previous children's stories), *report* and *confess* can take embedded clause complements (necessary in WSM constructions), but they carry additional meaning that adds complexity to the story. For example, *confessing* something is a different, more abstract event than simply *saying* something. In these experiments, the six embedded clause verbs *hire, avoid, scold, welcome, invite,* and *thank* were selected because one can plausibly either confess or report these actions, and because these verbs are more abstract than the verbs used in child-directed experiments (e.g. *pick, buy, steal, see, find, make* in Lutken et al. 2020). Another difference between these experiments and previous child experiments is that half of the questions were *how* questions and half were *why* questions, whereas in the child experiments, all of the questions were *how* questions. We included *why* questions to determine whether the WSM-like phenomenon was specific to *how* questions since *how* is an attested scope marker in languages like Polish (Stepanov 2000).

Regarding story design, the prominence of the matrix and embedded clause "events" was controlled. In the story shown in (4), for example, a comparable amount of time is spent discussing the *inviting* event and the *reporting* event.

(4) Sample Context Story for an Unambiguous Question:
The local Boy Scouts are having their jamboree. Arthur was given the job of inviting someone to teach knot-tying and someone to teach fire-building. He's excited about who he's asked to do fire-building, but he's concerned that the knot-tying expert might not show up. He thinks about reporting his concerns to the director but decides to wait just a bit longer. The director sends around an email asking everyone to confirm who they have invited to speak. Arthur goes ahead and reports that he's invited the fire-building expert, but decides he'd better not mention the knot-tying expert yet in case he doesn't work out.
*Q: Why did Arthur report who he invited?*

Each of the four possible interpretations of the question has an answer provided in the text. These include the correct answer (5a, because the director asked everyone to confirm who they invited); an incorrect matrix reading, indicating a correct interpretation of the question, but misremembering the correct response (5b, he is concerned the knot tying expert might not show up); an embedded question answer, indicating a response only to the embedded clause question (he invited a knot-tying expert and a fire-building expert); and a WSM response indicating the participant gave the *wh*-relativizer, *who,* scope over the entire question and answered who he reported he invited (the fire-building expert).

(5) Answer types for the unambiguous question *Why did Arthur report who he invited?*
a. Correct Matrix: Because the director asked everyone to confirm who they invited.
b. Incorrect Matrix: Because he was concerned the knot-tying expert might not show up.
c. Embedded Question: He invited the knot-tying expert and the fire-building expert.
d. WSM: He reported he invited the fire-building expert.

After half of the stories, we asked an ambiguous questions that lacked a *wh*-relativizer and so could be interpreted as a matrix or an embedded long distance (LD) question. These ambiguous questions served as a control condition and as a measure of how often participants chose the matrix or LD interpretation of biclausal sentences. An example of a story giving context followed by an ambiguous question is given in (6).

(6) Sample Context Story for an Ambiguous Question:
Jeremy is in medical school. He bumps into his old friend Lucy who invites him to lunch. He tells her he has to study and can't go. After Lucy leaves, Jeremy decides to study at Starbucks. He's just gotten his latte, when he sees Lucy! He really doesn't have time to chat, so he tries

to hide in the men's room to avoid her. The men's room is occupied, so he sneaks out the side door. Jeremy decides he'd better explain to Lucy that he was just grabbing a latte. He tries calling her, but she doesn't pick up. Jeremy ends up sending her an email apologizing. He explains that he is cramming for final exams and when he saw her at Starbucks, he slipped out the side door. He invites Lucy to lunch during winter break.
*Q: How did Jeremy confess he avoided Lucy?*

The question is ambiguous in that one could answer with how Jeremy did the confessing OR how Jeremy did the avoiding. The four answer types for the ambiguous story are given in (7):

(7) Answer types for the ambiguous question *How did Jeremy confess he avoided Lucy?*:
a. Correct Matrix: He emailed her.
b. Incorrect Matrix: He called her.
c. Correct Embedded: By leaving out the side door.
d. Incorrect Embedded: By hiding in the men's room.

*How* and *why* questions appeared as ambiguous and unambiguous questions equally often yielding the 4 types of trials shown in (8):

(8) Four Trial Types:
(a) Ambiguous *how* question: How did Jeremy confess he avoided Lucy
(b) Unambiguous *how* question: How did Gabrielle confess who she invited?
(c) Ambiguous *why* question: Why did Brenda report she invited Paul?
(d) Unambiguous *why* question: Why did Celia confess who she avoided?

Within an experiment, all participants read the same stories followed by the same questions. The order of the trials was pseudorandomized with the restrictions that: (1) no more than 3 trials with the same question word appeared in a row; (2) no more than 4 trials with the same matrix verb appeared in a row; (3) no more than 6 ambiguous or unambiguous question trials appeared in a row; and (4) no more than two trials with the same embedded clause question verb appeared in a row. Catch scenarios were distributed every 4-7 trials and were included in restrictions (2) and (4) since they used the same matrix and embedded clause verbs.

### 2.3 Procedure

In all three experiments, participants read a story and then answered a question about the story. The procedure varied slightly between the pilot and the experiments. The web-based experimental platform FindingFive (FindingFive Team 2019) presented stimuli and recorded participants' selections and reaction times in the experiments.

#### 2.3.1 Pilot Experiment

We first performed a pilot experiment designed to determine the efficacy of our story stimuli. There were two forms of the experiment, each of which contained 24 story-question pairs. Twelve participants did each form. The format of a trial was as follows. First, a story appeared on the screen. Once the participant finished reading the story, they hit a button, and the story disappeared and was replaced by an unambiguous or ambiguous question and a simple content question. The question remained on the screen while the participant typed their answer. Participants were free to type whatever they wished. If more than half of participants missed the content question, gave responses to our target questions that had not been mentioned in the story, or said they did not remember what happened, that story was either eliminated or rewritten for Experiments 1 and 2. Most of participants' responses (86%) corresponded to the types given in (5) for unambiguous questions and in (7) for ambiguous questions.

#### 2.3.2 Experiment 1

Experiment 1 differed from the pilot in the following ways. First, rather than typing their answer, participants chose their answer(s) from a list. Second, Experiment 1 included a working memory (WM) task. Third, Experiment 1 had more trials than the pilot experiment: 18 ambiguous and 18 unambiguous. Trials were balanced for ambiguous/unambiguous questions, *why/how* questions, and *confess/report* as the matrix verb. Based on our pilot experiment, 36 trials was the maximum number of stories participants could plausibly read and respond to in a reasonable time. Thus, although ambiguous and unambiguous *how* and *why* questions appeared equally often with the matrix verbs *confess* and *report*, the embedded clause verbs were not perfectly balanced among the 6 types of trials. Lastly, rather than have content questions after each trial, there were 6 simple "catch" trials that were designed to test whether participants were paying attention. The format of all trials was as follows. First, participants read a story. When they were done, they clicked a button and a digit string was displayed until the participant pressed a button or 3 seconds had elapsed, whichever came first. Next, the question appeared. Once participants had finished reading the question, they clicked a button, and the question disappeared and was replaced with the 4 possible answers to the question (see 5a-d, 7a-d). After participants answered the question, they selected which of two strings of numbers they had previously seen. The distractor string differed from the target string by reversing the position of two digits. One third of digit strings were two digits long, one third were 5 digits long, and one third were 7 digits long.

### 2.3.3 Experiment 2

The procedure for Experiment 2 was the same as Experiment 1 except there was no WM task.

## 3 Results

### 3.1 Pilot Experiment

The primary purpose of the pilot was to determine how many trials to include in the actual experiments, and to eliminate any stories that were too confusing. However, the results of this pilot were interesting in themselves. Of the 96 responses to unambiguous questions (12 participants x 8 uncut unambiguous trials), 67% (64/96) were correct. Probably reflecting the fact that the question remained on the screen while participants typed their answers, the vast majority of participants' responses (84.4%) were answers to the matrix verb question. Despite this, 4% (4/96) of the time participants gave WSM answers (e.g. saying who Arthur invited) and 3% (3/96) of the time participants gave island extraction answers (e.g. saying why Arthur invited people). [1]

### 3.2 Experiment 1

Because of space limitations, we only report the results for the first choice that participants selected for unambiguous questions.[2] Statistical analyses of participants' first choice responses were performed using JASP version 0.17.2 statistical software (JASP Team, 2023). Overall, participants chose the WSM option 13.5% (114/846) of the time for unambiguous questions. A 2 (matrix verb = *confess/report*) X 2 (*how/why*) Bayesian ANOVA with participant as a random variable revealed that the best fitting model ($BF_{10} = 64.68$) included main effects of both matrix verb and *wh*-word, and an interaction between the two factors (see Figure 1). Participants chose the WSM option more often for *how* than *why* questions (15.9% vs. 10.5%, respectively, $BF_{inc} = 7.548$), and for *report* than *confess* questions (16.2% vs. 10.2%, respectively, $BF_{inc} = 12.29$). As shown in Figure 1, the interaction ($BF_{inc} = 13.44$) was due to participants being more likely to give WSM responses to *how* questions with the matrix verb *report* (22.1%) than the other three types of questions.

[1]We did not include island extractions responses as answer choices in Experiments 2 and 3 because they were not the primary focus of this study.

[2]Because there are two possible correct answers for ambiguous questions, for both ambiguous and unambiguous trials, participants were allowed to choose more than one response. FindingFive time-stamped when participants selected options, and we used these time stamps to determine the order of selections.

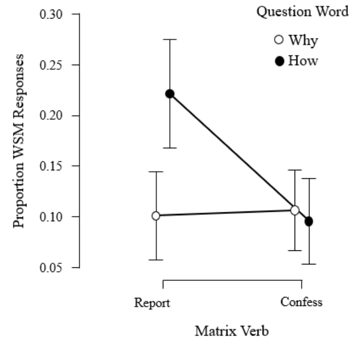


Figure 1: Interaction between Matrix Verb and *Wh*-word. Error bars are 95% credible intervals.

We next investigated the effect of the embedded verb on WSM responses via two Bayesian ANOVAs.[3] A 2 (*how/why*) x 6 (embedded verb) Bayesian ANOVA with participant as a random variable revealed that in the best fitting model (BF10 = 83124), there was decisive evidence of main effects of *wh*-word (BFinc = 3502.409, see above) and embedded verb (BFinc = 13064.09, see Figure 2A), which was due to relatively more WSM responses to *avoid*, *scold* and *welcome* and few WSM responses to *thank*, and an interaction between the two (BF inc = 901.76), which was largely due to there being more WSM responses for *how* questions with *scold* and *welcome* than *why* questions with *scold* and *welcome* (see Figure 2B). A 2 (matrix verb) x 5 (embedded verb) Bayesian ANOVA with participant as a random variable revealed that in the best fitting model ($BF_{10}$ = 49607, there was decisive evidence of main effects of matrix verb ($BF_{inc}$ = 19466, see above) and embedded verb ($BF_{inc}$ = 20935, see Figure 3A) and an interaction between the two ($BF_{inc}$ = 74613), which was largely due to there being more WSM responses for *report* questions with *scold* and *welcome* and fewer WSM responses for *report* questions with *invite* (Figure 3B).

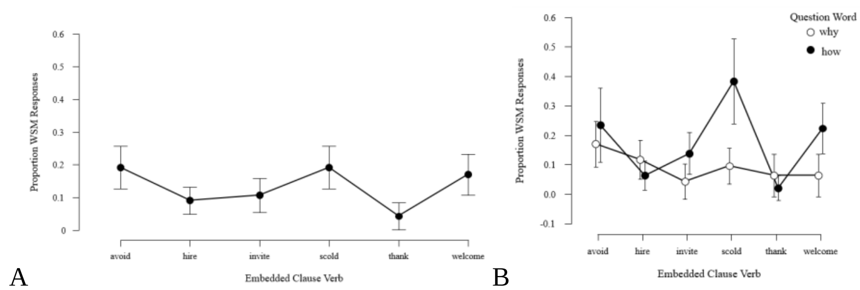


Figure 2: (A) Main Effect of Embedded Verb. (B) Interaction between *wh*-word and embedded verb. Error bars represent 95% confidence intervals.

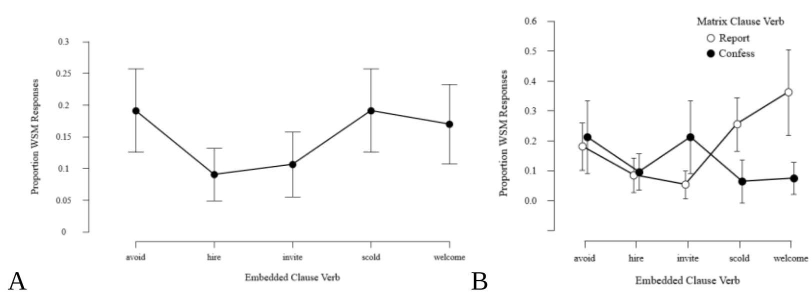


Figure 3: (A) Main effect of embedded verb. (B) Interaction between matrix and embedded verb. Error Bars represent 95% Confidence Intervals.

[3]In both Experiment 1 and 2, we were unable to conduct a single wh-word x matrix verb x embedded verb ANOVA because not every embedded verb appeared with each combination of *wh*-word and matrix verb.

A Bayesian ANOVA of WSM responses with story as the independent variable and participant as the random variable revealed decisive evidence of an effect of story ($BF_{10} = 5.6\times10^{+8}$), with the top 2 stories accounting for over 30% of all of the WSM errors and the bottom half of the stories accounting for only 20% of the WSM errors. A Bayesian ANOVA of WSM responses with participant as the independent variable and story as the random variable revealed decisive evidence of an effect of participant ($BF_{10} = 5.6\times10^{+7}$), with 11 participants giving no WSM responses, 19 participants giving 2 or 3 WSM responses, 16 participants giving between 4 and 7 WSM responses, and one participant giving WSM responses half the time. A Bayesian ANOVA of WSM responses with WM digit length as an independent variable and participant as a random variable revealed that the best fitting model was the null model, indicating that difficulty of the WM task did not affect whether participants gave a WSM response.

## 4 Experiment 2

The results from Experiment 2 were similar to those from Experiment 1. Overall, participants in Experiment 2 chose the WSM option 9.3% (40/432) of the time for unambiguous questions. As was the case with Experiment 1, a 2 (matrix verb = confess/report) X 2 (*how/why*) Bayesian ANOVA with participant as a variable revealed that the best fitting model ($BF_{10} = 110938$) included main effects of both matrix verb and *wh*-word, and an interaction between the two factors. Participants chose the WSM option more often for *how* than *why* questions (14.4% vs. 4.2%, respectively, $BF_{inc} = 4080$) and for *report* than *confess* questions (13.4% vs. 5.1%, respectively, $BF_{inc} = 335$). Like in Experiment 1, the interaction ($BF_{inc} = 293.780$) was due to participants being more likely to give WSM responses to *how* questions with the matrix verb *report* (21.7%) than the other three types of questions (see Figure 4).

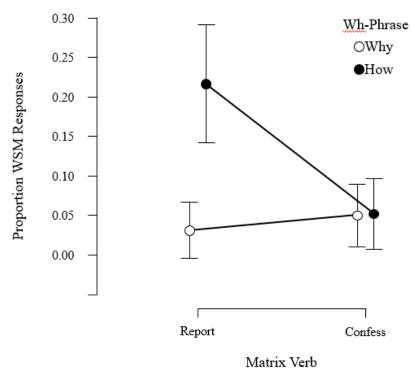


Figure 4: Interaction between Matrix Verb and *Wh*-word. Error Bars: 95% Confidence Intervals.

For Experiment 2, the 2 (*how/why*) x 6 (embedded verb) Bayesian ANOVA revealed that the best fitting model ($BF_{10} = 228$) had a main effect of *wh*-word, with participants giving more WSM responses to *how* questions than *why* questions (BFinc = 131). However, unlike in Experiment 1, there was no main effect of embedded verb, nor did embedded verb interact with *wh*-word. As was the case in Experiment 1, the 2 (matrix verb) x 5 (embedded verb) Bayesian ANOVA revealed that, in the best fitting model ($BF_{10} = 127$), there was evidence of main effects of matrix verb ($BF_{inc} = 67$, see Figure 5A) and embedded verb ($BF_{inc} = 16$) and an interaction between the two (BF inc = 11), which was largely due to there being more WSM responses for *report* questions with *welcome* (see Figure 5B).

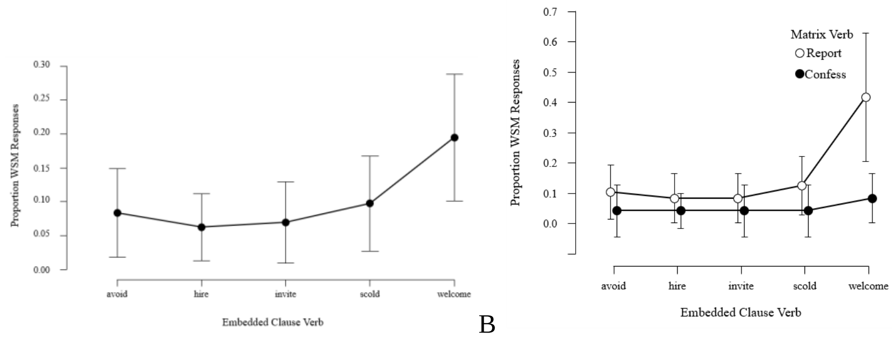


Figure 5: (A) Main effect of embedded verb. (B) Interaction between Matrix verb and embedded verb. Error bars represent 95% confidence Intervals.

A Bayesian ANOVA of WSM responses with story as the independent variable and participant as the random variable revealed decisive evidence of an effect of story ($BF_{10}$ = 7241), with the top three stories accounting for half of the WSM errors and the bottom two-thirds accounting for less than a third of the WSM errors. A Bayesian ANOVA of WSM responses with participant as the independent variable and story as the random variable revealed decisive evidence of an effect of participant ($BF_{10} = 2.757\times10^{+7}$), with about a third (9/24) giving no WSM responses, half (12/24) giving between 1 and 3 WSM responses, and 13% (3/24) giving 6 to 8 WSM responses.

The results from Experiment 1 and 2 were similar. In both experiments, participants usually answered unambiguous questions correctly. When they erroneously gave WSM errors, they did so in similar circumstances. In both experiments, they were more likely to incorrectly select the WSM answer for *how* questions than *why* questions, and for questions that had the matrix verb *report* than those that had the matrix verb *confess*. In both experiments, participants were more than twice as likely to give WSM responses to *how* questions with the matrix verb *report* (e.g. *How did Bob report who he avoided?*) than the other three types of unambiguous questions. In both experiments, the rate of WSM errors varied considerably by story (range 0%-40%), and specific stories tended to elicit similar rates of WSM errors in the two experiments ($r$ = .70628, $p$ = .00105). Lastly, in both experiments, there was considerable variability in the rate at which participants gave WSM responses, although in both experiments, most participants gave at least one WSM error, and in neither experiment did any participant give WSM responses more than half of the time.

There were, however, some differences between the two Experiments. As can be seen in Table 1, even though the stories were identical in Experiments 1 and 2, participants' responses were significantly different ($X^2$ (3, N = 1278) = 16.98, $p$ = .0007). Overall, participants were significantly less accurate ($X^2$ (1, N = 1278) = 7.51, $p$ = .006), and made more WSM errors ($X^2$ (1, N = 1278) = 4.80, $p$ = .029) in Experiment 1 than in Experiment 2. We suspect that the reason participants in Experiment 1 made more errors – and specifically more WSM errors – is that the participants in Experiment 1 had to do a working memory task and doing so overtaxed their processors and caused them to make more "child-like" WSM errors. Notably, in both experiments, adults made each type of error at rates that are comparable to children's rates in simpler studies designed for children. Specifically, the children in Lutken et al. (2020) gave correct matrix responses 53% of the time, WSM responses 16% of the time, incorrect matrix responses 6% of the time, and embedded question responses 4% of the time.

| Response Type | Exp 1 (N = 846) | Exp 2 (N = 432) |
|---|---|---|
| **Correct** (e.g. *The director asked everyone to confirm their invitations*) | 72.1% | 79.2% |
| WSM (e.g. *He reported inviting the fire building expert*) | 13.5% | 9.3% |
| **Incorrect Matrix** (e.g. *He was concerned about the knot-tying expert*) | 8.9% | 10.8% |
| **Embedded Question** (e.g. *He invited a fire building expert and knot-tying expert*) | 5.5% | 1.6% |

Table 1: Distribution of Responses to Unambiguous Questions in Experiments 1 and 2.

## 5 Discussion

By investigating how English-speaking adults interpret complex *wh*-questions with multiple clauses when their processing mechanisms are overtaxed, we hoped to shed light on why English-speaking children sometimes interpret these questions as if they were WSM questions. One possibility is that children make WSM errors because they have immature competence and consequently treat matrix *wh*-phrases as scope markers and *wh*-relativizers as contentful *wh*-words which they should answer. An alternative explanation is that children make these errors as the result of immature processing mechanisms which are over-taxed. If the immature competence account is correct, adults should never make WSM errors because nothing should force adults to revert to an alternate grammar. Under the immature processing account, on the other hand, English-speaking adults might sometimes produce WSM responses if their language processors are suitably overtaxed. As elaborated below, taken as a whole, we believe the results of our adult experiments support the immature processing account of English-speaking children's WSM errors.

To begin with, question word, matrix verb, and embedded clause verb all affected the rate of WSM responses. This is surprising under a competence account. One might plausibly argue that a competence account would predict more WSM errors with *how* than *why* because *how* is an attested scope-marker (Stepanov 2000), while *why* is not. However, this should apply equally with all matrix and embedded clause verbs, which is not what we found. Furthermore, it is unclear why WSM responses should occur more with *report* than *confess* or show variation among embedded clause verbs. On the other hand, a processing explanation predicts this variability because some verbs and question words or combinations of the two might tax processing mechanisms more than others. These WSM errors also do not seem to be the result of random guessing. If participants were simply choosing an answer at random, we would expect either even distribution across all response types or at least even distribution across incorrect responses, but this is not what we find. Furthermore, random guessing should not be affected by *wh*-phrase, matrix verb, or embedded clause verb and yet these affect the rate of participants' WSM responses. Thus, these WSM responses cannot be ascribed to participants making random guesses.

Furthermore, specific stories elicit more WSM errors than others. A competence account would have to posit a *grammatically*-based reason for why certain stories allow adults to use a WSM grammar while others would not. Such a reason is not apparent. However, under a processing account, stories and questions that are harder (for whatever reason) should elicit more errors than stories and questions that are simpler. The same stories that elicited WSM errors in Experiment 1 did so in Experiment 2, suggesting there is something more difficult about particular stories. We examined whether the number of words, clauses, or sentences in stories were related to the number of WSM errors but found no effect. We also investigated whether order of trial played a role, since participants might be more overtaxed toward the end of the experiment but, again, found no effect. While what makes certain stories and questions more difficult is as yet unclear, a processing account predicts such differences will exist, and a competence account has difficulty explaining them.

Although we were somewhat surprised that there was no clear relationship between the difficulty of the WM task and adults' error rate in Experiment 1, one possibility is that the linguistic task was already demanding enough that the WM task had little effect.[4] That there were significantly more errors overall and significantly more WSM errors in Experiment 1 than in Experiment 2 suggests that the presence of the WM task in Experiment 1 taxed the processing abilities of the participants in Experiment 1. This result, again, is difficult to explain under a competence account, yet is predicted under a processing account since greater processing load yields more errors.

We also found variation in performance within individual participants. Under a competence explanation, it is possible that some English-speaking adults might allow WSM. However, instead of a few participants making most of the WSM errors as a competence account predicts, over half of the participants made at least one WSM error and no individual made WSM errors more than half the time. This pattern is expected under a processing explanation since these errors are essentially the result of finding the story and/or the question difficult to process and making a mistake.

Finally, the pattern of adult errors resembles the pattern of children's errors found in previous

[4]We cannot rule out the alternative possibility that we failed to find an effect of the WM task on error rates because our working memory task taps different processing resources than our linguistic task.

studies (Lutken et al. 2020, Lutken and Legendre 2021). This is consistent with children and adults responding similarly when presented with tasks that tax them to a similar extent. If adults make WSM errors when they are given an appropriately difficult task (i.e., when the task is designed for adults and not for children or when they are simultaneously doing a WM task), then it stands to reason children should do so as well.

## 6 Conclusion

Our question was simple: if we tax English-speaking adults' language processors, do they make WSM errors just like children do? Our answer is equally simple: they do. While Lutken et al. (2020) and Lutken and Legendre's (2021) adult controls performed perfectly on a task designed for children, we found that when we gave adults tasks that were linguistically similar, but more taxing, they made exactly the same types of errors children make, and they did so at similar rates to children. The fact that both adults and children make WSM errors suggests that WSM errors reflect a processing failure that both children and adults are susceptible to when their processing abilities are taxed. Thus, one cannot take children's WSM comprehension errors as proof that they have WSM-grammars. Our findings also have more general implications for this type of work. First, a cautionary tale: children's performance on psycholinguistic tasks shouldn't be taken as a pure indicator of their competence any more than adults' occasional failure to understand center-embedded sentences like *The cat the dog chased died* should be taken to indicate adults' grammars do not have embedding. Second, comparing adult and child performance on tasks designed for children is fraught. What looks like adults having greater syntactic competence may simply reflect them having greater processing abilities.

## References

de Villiers, Jill., Kotfila, Jessica., and Klein, Madeline. 2019. Parsing, pragmatics, and representation: Children's comprehension of two-clause questions. In *Three Streams of Generative Language Acquisition Research*, ed. T. Ionin and M. Rispoli, 85–105. Amsterdam: John Benjamins.

de Villiers, Jill and Thomas Roeper. 1995. Relative clauses are barriers to *wh*-movement for young children. *Journal of Child Language* 22:389–404.

FindingFive Team. 2019. FindingFive: A web platform for creating, running, and managing your studies in one place. FindingFive Corporation, NJ, USA

JASP Team. 2023. JASP (Version 0.17. 2) [Computer software]

Lutken, C. Jane, Geraldine Legendre, and Akira Omaki. 2020. Syntactic Creativity in Children's Wh-Question. *Cognitive Science* 44:E12849.

Lutken, C. Jane and Geraldine Legendre. 2021. A Cross-linguistic comparison in children's interpretations of complex questions. Paper presented at Linguistics Society of America 96, Virtual Conference.

Lutz, Uli., Gereon Müller, and Arnim von Stechow. Eds. 2000. *Wh-scope marking*. Philiadelphia, PA: John Benjamins.

Omaki, Akira and Jeff Lidz. 2015. Linking parser development to acquisition of syntactic knowledge. *Language Acquisition* 22:158–192.

Snedeker, Jesse. 2013. Children's sentence processing. In *Sentence Processing*, ed. R. P. G. von Gompel, 189–220. New York: Psychology Press.

Snyder, William. 2007. *Child Language: The Parametric Approach*. New York: Oxford University Press.

Stepanov, Arthur. 2000. WH-scope marking in Slavic. *Studia Linguistica* 54:1-40.

Thornton, Rosalind, and Stephen Crain. 1994. Successive cyclic movement. In *Language Acquisition Studies in Generative Grammar,* ed. T. Hoekstra and B. Schwartz, 215–252. Philadelphia, PA: John Benjamins.

Yang, Charles. 2002. *Knowledge and Learning in Natural Language*. New York: Oxford University Press.

Rutgers Center for Cognitive Science
Rutgers University: The State University of New Jersey
Piscataway, NJ 08854
*c.jane.lutken@rutgers.edu*
*karin.stromswold@rutgers.edu*