THE RELATIONSHIP BETWEEN CHILDREN'S METACOGNITIVE JUDGMENTS OF KNOWLEDGE AND VERBAL DISFLUENCY

by

Eloise West

B.A., Barnard College of Columbia University, 2020

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT OF THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Psychology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2022

© Eloise West, 2022

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

The relationship between children's metacognitive judgments of knowledge and verbal disfluency

submitted by Eloise West in partial fulfilment of the requirements for

the degree of Master of Arts

in Psychology

Examining Committee:

Darko Odic, Associate Professor, Psychology, UBC

Supervisor

Janet F. Werker, University Killam Professor, Psychology, UBC

Supervisory Committee Member

Susan Birch, Associate Professor, Psychology, UBC

Supervisory Committee Member

Abstract:

When we [uhh] have everyday conversations, our speech is [um] littered with [like] spontaneous pauses and interjections known as “verbal disfluencies”. In adults, verbal disfluencies are associated with a speaker’s certainty or knowledge level in both speech production and speech comprehension. That is, adults rate both their own and others’ confidence lower when they produce more verbal disfluency. Little work has explored whether and when children’s verbal disfluency correlates with their own internal sense of confidence. Given that young children struggle with explicit ratings of their own confidence, these implicit cues may provide researchers a window into children’s metacognitive awareness. This study examines the association between verbal disfluency and confidence in 5–8-year-olds’ (N=60) naturally produced speech. Children answered fact-based questions about animals and performed numerical comparisons. Then, they rated their confidence about these answers in a forced-choice metacognitive judgment paradigm. We examine the association between verbal disfluency and the accuracy of children’s responses, as well as these explicit ratings of metacognitive confidence, showing that even our youngest children reliably produce more verbal disfluencies when they answer incorrectly, and when they feel less confident. Moreover, children’s verbal disfluencies predicted the accuracy of their response over and above their explicit ratings of confidence, suggesting that future work should consider examining verbal disfluency as a measure of children’s metacognition.

Lay Summary:

When adults feel less confident, we use more pauses, fillers like “ums” and “uhs”, and hedging language like “I think”—speech cues known as “verbal disfluencies”. Here, we ask whether young children (ages 5-8) use verbal disfluencies like adults. Kids answered questions about animals and number and rated their confidence in their answer. We found that children used more disfluency when they answered questions incorrectly, and when they indicated feeling less confident. This suggests that, like adults, children’s verbal disfluency conveys information about their internal feeling of confidence.

Preface:

This thesis is an original, unpublished work. The research was conducted at the University of British Columbia's Centre for Cognitive Development. I completed the experimental design, data analysis, and writing under the supervision of Dr. Darko Odic. The forced-choice metacognitive judgment paradigm was adapted from a task designed by Dr. Carolyn Baer.

The Behavioural Research Ethics Board at the University of British Columbia approved all work in this thesis (H21-00682 Kid Disfluency and Confidence).

Table of Contents

List of Tables

List of Figures

Acknowledgements

Thank you to my supervisor, Darko Odic, for all the insightful feedback and advice throughout this thesis. I am so grateful for your patience and support these past years.

This project would not have been possible without the help of Carolyn Baer, who designed the forced-choice metacognitive judgment paradigm after which this task was modeled. Thank you for your guidance in navigating this literature.

To the team of research assistants who volunteered their time to help with recruiting participants, running the study, and transcribing our data – Robyn Armstrong, Kaja Bakken, Alice Erchov, Tasmia Jahan, Steffi Lau, Aimee Lutrin, Kiana Rashidi, Isabella Shoettler, Eliscia Sinclair, Martyna Siwocha, Megan Toi, and Lisa Yu – I am forever grateful for your help. This project would not have been possible without you.

And of course – my sincerest gratitude for all the children and families who volunteered their time to participate in our research.

Introduction

[uhhh] Real life speech is [um] littered with [like] spontaneous pauses and interjections. Such interjections are known as "verbal disfluencies". These terms are remarkably common in conversation – estimates range between 2 and 26 disfluencies per 100 words (Fox Tree, 1995), and disfluencies exist cross-culturally in many of the world's languages (Amiridze et al., 2010; Crible, 2018; Tian et al., 2017), across age, gender, and socio-economic demographics (Bortfeld et al., 2001; DeJoy & Gregory, 1985; Jansson-Verkasalo et al., 2021; Kools & Berryman, 1971; Laserna et al., 2014; Tottie, 2011, 2014; Yairi & Clifton, 1972).

Despite their ubiquity, these terms often carry negative connotations. For example, although disfluencies are common across ages and genders, they are associated with gendered stereotypes (e.g., the Californian English "Valley Girl" accent; Hinton et al., 1987; Preston, 1986). It is no surprise then, that many popular journals offer advice or training on how to reduce the number of disfluencies in speech for self-improvement (Dizik, 2016; Mele, 2017; Zandan, 2018).

Though some seek to purge natural speech of all disfluencies, psychology and linguistic research shows that they serve a valuable role in communication. Disfluencies are, for example, associated with one's metacognitive confidence: English-speaking adults produce more disfluencies when they explicitly report feeling less certain (Smith & Clark, 1993), and they rate other speakers as being less knowledgeable when they produce more verbal disfluencies (Brennan & Williams, 1995). Verbal disfluencies, therefore, can help inform our understanding of others' minds.

While these cues facilitate the social dynamics of communication, no one explicitly teaches us to use them—if anything, they discourage us from producing them—or tells us that

they signal a speaker's level of confidence. To understand how we learn the association between confidence and verbal disfluency, we must turn to the developmental story: when, and how, do children learn the communicative role of verbal disfluency?

This thesis explores when and how children use disfluency to indicate their own certainty or confidence. We adapt Smith & Clark's (1993) task to investigate the relationship between young children's (ages 5-8 years) verbal disfluency in answering a range of fact-based and perceptual questions, and their explicit metacognitive ratings of certainty in their responses. The central question of this thesis is when children's *production* of disfluent speech predicts their own subjective sense of certainty in their decisions. By treating disfluencies as a genuine part of the communicative role of speech, the results have implications for theories of language learning (e.g., how do children learn the relationship between disfluency and confidence), and metacognition (e.g., what do children's disfluencies reveal about their early representations of confidence and certainty).

I will first begin with a broad review of verbal disfluency: *what* are disfluencies and *when* do we produce them? Then I will review a debate regarding whether disfluency is an intentional signal from to speaker, or an incidental consequence of taking time to think, and discuss the implications of these approaches for children's developing use of disfluency. Finally, I will propose that young children's verbal disfluencies may not always track their explicit reports about their confidence, as their explicit metacognitive judgments are influenced by a range of task-demands on performance, before describing our current experiment. I will then focus on answering four empirical questions: (1) Do English-speaking children produce verbal disfluency and, if so, what *kinds*; (2) Do children produce more disfluency when they are less accurate; (3)

Do children produce more disfluency when they are less confident in their answers; and (4) Are disfluencies a *better* predictor of children's accuracy than their explicit confidence judgements?

What are verbal disfluencies?

Verbal disfluencies are the extralinguistic pauses and interjections in our speech: the "ums", "uhs", and (in some contexts) "likes", for example. Some linguists provide a taxonomy of differential disfluency types, often arguing that each term carries unique semantic meaning: distinguishing between fillers (e.g., "um", "uh", "er"; with some differentiating *between* individual filler terms like "um" and "uh"; Clark & Fox Tree, 2002), self-repairs, including repeats (e.g., "just on the left-left side") and restarts (e.g.,"imme-just below the left side"), editing expressions (e.g., "I mean", "sorry"; Bortfeld et al., 2001), and hedges (e.g., "I guess", "like"; Fox Tree, 2006; Smith & Clark, 1993). Throughout this paper, however, I will use "verbal disfluency" as an umbrella term to refer to all these different parts of speech.

Verbal disfluencies exist in many languages across the globe. While individual disfluency *terms* differ across languages, they share a common function, cross-linguistically, in organizing discourse structure (Amiridze et al., 2010; Hayashi & Yoon, 2006). No matter their token form, disfluencies rely on similar pragmatic markers across languages: phonological devices like sound stretch (e.g., English and Ilokano; Streeck, 1996) or rising pitch and intonation (e.g., English and Dutch; Krahmer & Swerts, 2005; Smith & Clark, 1993), pauses or silence (Stivers et al., 2009), self-repair (e.g., English, German, & Hebrew; Fox et al., 2010), and interjections. Interjections, cross-linguistically, share several common forms: non-lexical but conventionalized sounds, or fillers (e.g., English "uh" and Japanese "eto"; Clark & Fox Tree, 2002; Watanabe et al., 2008), as well as lexical items like demonstratives (e.g., English "thee" and Japanese "ano"; Arnold et al., 2007; Watanabe et al., 2008), discourse markers (e.g., English

"so" and French "hein"; Crible et al., 2017) and placeholder fillers (e.g., English "whatchamacallit" and Russian "eti"; (Podlesskaya, 2010).

Notably, while disfluencies share common forms across languages, this does not necessitate that they communicate the same information. For example, in English, fillers often precede a pause, suggesting memory retrieval difficulty (Clark & Fox Tree, 2002; Smith & Clark, 1993). In Mandarin Chinese, fillers can be used to signal the syntactic category of the next word, while Japanese uses different fillers to communicate varying degrees of interaction between interlocutors (Tian et al., 2017). Thus, despite the cross-linguistic similarities in verbal disfluency, each individual language warrants its own investigation, to determine cross-cultural differences in *what* specific disfluency tokens communicate, and how children might learn these pairings.

When do we produce verbal disfluencies, and what do they communicate?

To appreciate the communicative role of verbal disfluency, one must consider the complex social and cognitive processes that underlie everyday conversation. Communication is intrinsically a collaborative social process where conversational partners must coordinate to achieve common ground, both in the *process* and the *content* of conversation (Clark & Brennan, 1991; Grice, 1989; Schober & Brennan, 2003). One must coordinate the *process* of conversation by appropriately synchronizing turn-taking between partners, and the *content* of conversation by coordinating shared knowledge, beliefs, and assumptions, for example, assuming a communal understanding of word meaning (Clark & Brennan, 1991).

Disfluencies can facilitate the coordination of *both* the process and the content of communication. They coordinate the *process* of communication in providing a signal to

conversational turns or discourse structure (Beňuš et al., 2011; Bortfeld et al., 2001; Clark, 2002; Swerts, 1998; Walker et al., 2014) and in guiding speech parsing (Bailey & Ferreira, 2003; Ferreira & Bailey, 2004). Disfluencies can coordinate the *content* of conversation, too, in serving as cues to novel objects of reference (i.e., in eye-tracking experiments, "put theee, uh", but not "put the", guides attention toward discourse-new objects in the scene; Arnold et al., 2003, 2004; Barr, 2001; Barr & Seyfeddinipur, 2010), and impacting word recognition (i.e., words following filled pauses are recognized faster; Corley & Hartsuiker, 2011; Fox Tree, 1995; Fox Tree & Schrock, 1999). Therefore, disfluencies impact both the temporal aspects of conversation, setting the time-course of turn-taking, as well as the actual content of conversation, cueing speakers to shared referents.

But beyond the role of coordinating the process and content of a conversation, disfluencies serve a social pragmatic role in communication. Disfluencies can inform our understanding of others' minds, cueing us to one's metacognitive monitoring processes, including how certain or confident an individual feels about their knowledge. This is in part evidenced by the relationship between difficulty and disfluency (e.g., syntactic complexity; Cook et al., 1974; Maclay & Osgood, 1959; Watanabe et al., 2008; conceptual complexity; Barr, 2003) – wherein it is assumed that one intrinsically feels less knowledgeable or confident in producing more complex utterances. Stronger evidence, and more relevant to the current study, comes from the association between verbal disfluencies and conscious, *explicit* metacognitive ratings of one's knowledge level. Adults produce more verbal disfluency when they report feeling less certain (Smith & Clark, 1993), and rate other speakers as being less knowledgeable when they produce more disfluency (Brennan & Williams, 1995). Indeed, several prominent theories of metacognition suggest that confidence is the self-perception of (dis)fluency: that we determine

our level of confidence, explicitly, by the subjective experience of cognitive fluency, or how easily the information is accessed from memory (Alter & Oppenheimer, 2009; Koriat, 1993).

In one study, adult participants were asked factual questions in a conversational setting—questions like "in which sport is the Stanley Cup awarded?". After answering each question, participants provided explicit metacognitive judgments of their "Feeling of Knowing", indicating the likelihood they would be able to recognize the correct answer on a scale of 1 ("absolutely sure I WON'T recognize") to 7 ("absolutely sure I WILL recognize") (Smith & Clark, 1993). Researchers transcribed their verbal responses, paying special attention to the duration of disfluencies like fillers and pauses, and looked at how these extralinguistic markers related to their explicit ratings of metacognition. The lower their feeling of knowing, the more often they marked their responses with fillers, self-talk, hedges, and rising intonation (Smith & Clark, 1993). In other words: we produce more disfluency when we feel less confident (or, perhaps, we feel less confident when we produce more disfluency).

Verbal disfluencies are not only associated with our *own* confidence in speech *production*, but they also relate to our perception of *others'* confidence, in speech *comprehension*. In a subsequent study, Brennan and Williams (1995) first replicated the method and pattern of results in Smith & Clark (1993) wherein participants produced more disfluency on trials where they indicated a lower feeling of knowing. Then, they used the recordings obtained from these participants in a second experiment, examining this effect in speech comprehension. A new set of adult participants listened to the recorded responses of participants from experiment one and provided judgments of their "Feeling of Another's Knowing", indicating whether they believed the participant's response was the correct answer on a scale from 1 ("definitely incorrect") to 7 ("definitely correct"). Disfluent answers (but not non-answers, i.e., "I don't

know" responses) led to lower ratings of another's knowing (Brennan & Williams, 1995). Thus, we use disfluency to communicate about our own confidence, as well as to understand others' confidence.

Do we intentionally produce disfluency to express our confidence?

While we understand that disfluencies relate to confidence, both in speech production and speech comprehension, whether disfluency is an intentional signal from the speaker, or an incidental consequence of uncertainty, remains debated (Finlayson & Corley, 2012). Some argue that disfluency should be excluded from linguistic analysis and theory, as it reflects a *performance error* in applying linguistic knowledge to real-world speech (Chomsky, 1965). Indeed, disfluencies appear to be, at least in part, an error on the level of inhibition: individuals with attention-deficit/hyperactivity disorder (ADHD, combined subtype), a disorder associated with inhibitory control difficulties, produce more repetition and repair disfluencies than a non-ADHD control group (Engelhardt et al., 2010). However, other theories see disfluencies as genuine parts of language, arguing that fillers like "um" and "uh" should be considered actual words with differential semantic meaning (Clark & Fox Tree, 2002). As noted above, in some studies adults reliably use "uh" to signal brief delays and "um" longer ones, supporting a semantics-based account (Smith & Clark, 1993).

Independent of whether we consider disfluency a part of language or a performance error, it can still communicate the same metacognitive content. Even if a speaker does not intend to *produce* disfluency, this does not necessarily impact the listener's *perception* of disfluency. And even if we ourselves do not intend to produce disfluency, it may still correlate with our internal feeling of confidence. However, determining whether disfluency is an intentional signal from the

speaker has implications how we *acquire* disfluency. If disfluency reflects an error in performance, then it's production as a confidence signal need not be learned. But if we willfully interject "ums" and "uhs" to intentionally signal to others that we need time to think (Clark & Fox Tree, 2002), as we do any other word, then these must be learned like any other word, too.

Learning that disfluency refers to the state of another's mind – the speaker's subjective feeling of confidence – may prove more challenging than learning other words, however. While many nouns and verbs have visual referents, another's confidence – their subjective judgement of the probability of being correct – is not necessarily external or observable without first learning that, e.g., shrugging shoulders are an indicator of low confidence. Moreover, understanding a person's confidence requires some theory of mind, to understand that another's metacognitive monitoring operates over their own knowledge, which may be different than one's own. Children's acquisition of intentional disfluency – especially as related to their confidence – may, therefore, show protracted development relative to other aspects of language.

Disfluencies, of course, exist from the time we begin to use language – children as young two have been observed to produce verbal disfluency in various languages (DeJoy & Gregory, 1985; Jansson-Verkasalo et al., 2021; Yairi, 1981) – but looking at children's production of disfluency alone cannot reveal whether they understand them as communicating confidence. Children may produce disfluency without communicative intent: it is possible that children's use of disfluency is a natural consequence of speaking – a performance error – that with development is linked to explicit representations of metacognitive confidence. Or it may be that they are simply imitating or parroting those around them, and later learn the association to their metacognitive monitoring. Thus, to determine whether children understand disfluency to relate to

confidence, one must look at how it relates to their explicit ratings of their own and others' metacognition.

Many studies have investigated children's perception of disfluency in *others'* speech. Most support for an early understanding of the link between disfluency and confidence comes from the downstream behavioral and cognitive consequences of the perception of speaker confidence: when a speaker produces disfluency, and appears uncertain, this influences our trust in their opinion and the likelihood we will learn from them. For example, 2-year-olds selectively imitate fluent speakers relative to those who use disfluencies (Birch et al., 2010), and 3- to 4-year-olds selectively endorse novel object labels from more fluent speakers (White et al., 2020), providing converging evidence that young children perceive disfluent speakers to be less certain or reliable.

Some studies have more directly examined this relationship, asking children to explicitly rate others' knowledge or certainty. Pauses and response latency influence a host of children's social attributions: children assumed slower response times reflected the source of the speaker's knowledge, or the cognitive mechanisms underlying their answer—whether they were relying on memory retrieval, perception, or inference—as well as the complexity and accuracy of their knowledge (Richardson & Keil, 2022). Another study adapted the method of Brennan & Williams (1995) for young Dutch-speaking children (Krahmer & Swerts, 2005). Children's ratings of another's knowing similarly tracked the speaker's verbal disfluency: children rated others' knowledge lower when they produced more verbal disfluency, although their ratings of another's knowing were overall less accurate than adults'. Thus, both implicit and explicit measures suggest that young children understand disfluencies as communicating information about speaker certainty or confidence.

The *perception* of disfluency in others' speech, however, cannot discern whether the speaker intended to use these cues: recall that a performance error from the speaker can still communicate the same metacognitive content to the listener. To understand whether children use disfluency to communicate their own confidence, we must turn to their *production* of disfluency, and examine how it relates to their own feelings of confidence.

Surprisingly few studies have investigated how children's disfluency production relates to their explicit judgments of certainty (Hübscher et al., 2019; Krahmer & Swerts, 2005; Visser et al., 2014). One study asked Catalan-speaking preschoolers (3-5 year-olds) to identify occluded novel objects by touch and indicate how sure they felt about their answer using a 3-point verbal Likert scale ("very", "somewhat", or "not very"). While children produced more disfluency cues when identifying less familiar objects, they observed a dissociation between their metacognitive judgments and their facial and prosody cues: children often verbally reported that they felt certain, but used uncertain facial expressions and prosodic markers (Hübscher et al., 2019). The authors propose, therefore, that behavioral cues to uncertainty emerge earlier in development than lexical markers, and serve as bootstrapping devices for later sociopragmatic and metacognitive development (Hübscher et al., 2017; Hübscher & Prieto, 2019).

Another study adapted the method of Smith & Clark (1993) to be appropriate for young children (7-8 years old). Dutch-speaking children answered fact-based questions (e.g., "what is the color of peanut butter?") and provided explicit ratings of their metacognition via a Feeling of Knowing rating on a 5-point Likert scale with cartoon face analogs (i.e., 1 = frowning face, 5 = smiling face). Children exhibited the same general pattern of results as adults in Smith & Clark (1993): they produced more verbal disfluency on trials where they indicated lower feelings of knowing, however, this association was weaker than the one found among adult participants

(Krahmer & Swerts, 2005). Furthermore, children produced different disfluencies than adult participants. Adults primarily used fillers, pauses, high intonation and facial cues (raised eyebrow and funny faces) to mark their answers. Children's fillers, in contrast, showed no relationship with their FOK scores. Thus, while fillers may be the clearest disfluency cue to uncertainty in adults (Brennan & Williams, 1995; Smith & Clark, 1993), these may be less reliable cues in children.

While these studies provide preliminary evidence that children's disfluencies relate to their confidence, there are several reasons to question the generalizability of these results. First, given the aforementioned differences in the content communicated by individual disfluency terms (e.g., Mandarin Chinese, Japanese, and English fillers; Tian et al., 2017), it is important to examine this effect cross-linguistically. Second, one must consider how this association may shift with development. Krahmer & Swerts (2005) examine this effect in older children, 7-8 year-olds. By age 7, however, children are quite experienced language users. To determine whether this relationship is learned, one must probe for this phenomenon in younger kids. While Hübscher et al. (2019) examined this effect in preschoool-aged children (3-5 year-olds), they collapsed all hedges, self-talk, and vague language into one measure (lexical markers), and all fillers, pitch changes, and vowel elongations into one measure (prosodic markers). While it is possible that all lexical cues and all prosodic cues track confidence in the same way, some theories hold that these should be analyzed separately. Recall that some propose that individual fillers, for example, carry differential semantic meaning (Clark & Fox Tree, 2002). This theory implies that children must learn the semantic meaning of "um" and "uh" just as they do any other word – and would therefore use these variably like adults. However, while adults may understand "um" and "uh" to have differential semantic meaning (Clark & Fox Tree, 2002),

young children may not: 3- and 4-year-old children to do not use “um” and “uh” to reliably signal different degrees of disruption (Hudson Kam & Edwards, 2008). This suggests that children know the basic function of fillers, but do not yet differentiate between them. It is possible, however, that children begin with a nascent understanding of the relationship between certainty and disfluency that is enriched with development and later acquisition – wherein they learn to differentiate between terms’ use and function. It is essential, therefore, to examine children’s use of disfluency across a wide age range, and to consider how they may differentially use individual disfluency terms.

Finally, children have repeatedly been shown to be highly biased toward overconfidence when using traditional verbal report and Likert-scale metacognitive judgment tasks, often indicating certainty in answering questions incorrectly (Bayard et al., 2021; Destan & Roebers, 2015; Finn & Metcalfe, 2014; Kim et al., 2016; Lipko et al., 2009; van Loon et al., 2017), limiting the sensitivity with which we can establish a relationship between confidence judgements and certainty. This overconfidence bias may not reflect a lack of metacognitive monitoring, however: it is possible that the standard measures of metacognitive reasoning used with adults are too difficult for children of this age (Goupil & Kouider, 2019; Lyons & Ghetti, 2010). Instead, children’s overconfidence may emerge from a range of other social and cognitive pressures, unrelated to their metacognitive processing. One such theory is “wishful thinking”: that children judge their performance based on their idealized view of how they would like to be (Schneider, 1998). This might be a particularly strong demand in lab settings – where kids may feel extra pressure to signal to the experimenter that they are “smart”, or confident in their answers. This highlights the need for a metacognitive judgment task that maximizes children’s success – one that removes task demands, and the ability to respond overconfidently. By using

more developmentally appropriate measures of confidence that control for bias (e.g., Baer et al., 2018, 2021; Baer & Odic, 2019, 2020; Butterfield et al., 1988), we can get a finer picture of how disfluencies relate to confidence in even younger children.

While reducing the granularity of a Likert scale (e.g., from 7-points to 3-points; as in Hübscher et al., 2019) can make these tasks easier for children, it still allows for response biases (i.e., children can respond 3 ("very confident") on every trial). Indeed, the preschoolers in Hübscher et al. (2019) tended to overestimate their confidence, and therefore their metacognitive judgments did not track the accuracy of their responses, as observed in adults. Their behavioral cues to confidence – facial expression, gesture, and prosodic markers – in contrast, did track their accuracy: children used more uncertain expressions, posture, and prosody, when they answered incorrectly. This divergence may be because, while children's explicit metacognitive judgments are influenced by external task demands – like the "wishful" desire to appear competent (Schneider, 1998) – their disfluency may not be. For example, a follow-up to Krahmer & Swerts (2005) found that older children's (11-year-olds, but not 8-year-olds) explicit feeling of knowing ratings were sensitive to social context (collaborative vs. competitive), their verbal disfluency, in contrast, was not (Visser et al., 2014), suggesting this implicit measure is insensitive to social and task demands. It is possible that in using a response format that facilitates metacognitive judgments in younger children, one that does not allow them to respond "very confident" on every trial, we may no longer observe the dissociation between preschoolers' verbally reported confidence ratings and behavioral cues to certainty (Hübscher et al., 2019)

Alternatively, it may be that disfluency will track the accuracy of their responses better than any form of explicit judgment, even after adjusting the metacognitive judgment paradigm. It may be that disfluency is simply a better window into children's metacognition than their self-

reports and will track the accuracy of their responses better than the explicit judgments, across paradigms. To arbitrate between these accounts, again, it is essential to employ a metacognitive judgment task that gives children the best opportunity to demonstrate their metacognitive knowledge: one that removes task demands, like the ability to respond overconfidently. It is possible that, when given a task that facilitates accurate metacognitive judgments, we will find that disfluency and explicit judgments will both track the accuracy of their response

To summarize: verbal disfluencies have been consistently associated with explicit confidence judgements in adults, both in production and comprehension. But the developmental question of when children *produce* disfluencies to signal something about their confidence remains underexplored, especially in preschoolers, and typically confounds multiple types of disfluency markers and utilizes metacognitive measures that are prone to bias. By taking a more fine-grained approach at measuring both disfluencies *and* metacognition, we can better understand if even young children have a robust association between disfluency and their confidence judgements *and* the degree to which the overlap between the two might strengthen with age.

The present study.

In this thesis, I explore how children's verbal disfluency relate to their explicit metacognitive judgments of certainty and to the actual accuracy of their responses. Put differently, I examine whether produced disfluencies are an *implicit* measure of confidence (do they correlate with probability of success?) and how they relate to *explicit* measures of confidence. I adapt the method of Smith & Clark (1993) to be appropriate for young children, like Krahmer & Swerts (2005) – extending these investigations to English-speaking kids. I also

separate various *types* of disfluencies to examine whether children generate and use them in distinct ways.

I adapted the mode of confidence judgment: rather than asking children to estimate the likelihood they would recognize the correct answer, I gave them a forced-choice metacognitive judgment task that maximizes their success, and minimizes the social pressures and task demands that may underlie children's overconfident evaluations of their metacognition. A forced-choice task – wherein we ask children two questions, and then ask them which of the two was their better answer – does not allow kids to respond overconfidently, as they must reason about which of their answers was, relatively, *more* likely to be correct. In other words, they cannot respond "very certain" on every trial (Baer et al., 2021). In removing children's overconfidence bias, children's metacognitive judgments using a relative, forced-choice task may better track their accuracy than using a standard Likert scale. Indeed, in a range of tasks, children's relative judgments of confidence tracked their accuracy (Baer et al., 2018, 2021; Baer & Odic, 2019, 2020; Butterfield et al., 1988). Thus, this relative measure of confidence maximizes the possibility that children's disfluency will correlate with their metacognitive judgments and will allow us to further assess whether implicit or explicit cues provide a better measure of children's metacognition.

I also changed the questions children are asked: not only making them more accessible to children (e.g., "what do cows eat?" instead of "what is the capital of Chile?"), but also including various question *types* (numerical comparison, animal identification, and animal fact questions). Notably, some of these question types are open-ended questions, while others are binary choices, allowing me to also examine if disfluencies function differently for memory retrieval versus perceptual decision-making. I also varied the difficulty of the questions asked in effort to induce

a range of certainty and confidence in children (See Tables 10 and 11 in the Appendix for a list questions and answers, and accuracy rates).

Together, I focus on four primary questions, and additionally provide several exploratory analyses and suggestions for future work:

(1) Do English-speaking preschool-aged children produce verbal disfluencies and, if so, what kinds?

(2) Do disfluencies that children produce relate to the accuracy of their responses? In other words, could disfluencies be an *implicit* measure of confidence states? I can answer this question not just by examining general disfluency production across correct versus incorrect answers, but also further looking into whether specific *types* of disfluencies are better or worse markers of accuracy.

(3) Do disfluencies that children produce relate to the explicit confidence in their responses? This question is especially important in younger children, as previous work has suggested dissociations (Hübscher et al., 2019; Visser et al., 2014), but may have confounded metacognitive sensitivity for bias.

(4) Are disfluencies a *better* predictor of accuracy than explicit confidence judgements, even when controlling for bias? Answering this question would not only help future developmental psychology work methodologically (e.g., by allowing for the collection of metacognitive data in the absence of any explicit confidence judgements), but would also be relevant for many theories of metacognition that posit that the perception of (dis)fluency informs explicit feelings of confidence (e.g., Alter & Oppenheimer, 2009; Koriat, 1993).

Methods

Participants. Sixty children ages 5 to 8 years-old participated in the study, with 15 children in each age group (5;0 - 5;11, *M* = 5.42 years, 6;0 - 6;11, *M* = 6.40 years, 7;0 - 7;11, *M* = 7.43 years, 8;0 – 8;11, *M* = 8.51 years). We selected this age range as children ages 5-and-up can reliably report their metacognitive awareness. Given this study involved understanding and answering questions in English, participants were required to hear at least 50% English in their daily life – validated by parent report via an online form. An additional 13 participants were tested but were excluded from the analyses. We excluded 8 participants due to oversampling – we tested more participants than pre-registered – the first 8 participants were removed prior to analysis. Other reasons for exclusion include failing to complete the experiment (1), responding in a language other than English (1), experimenter or equipment error (e.g., the testing session was not recorded (2), or the participant's microphone quality interfered with the recording (1)).

Children were recruited from the Early Development Research Group database of volunteers from the Greater Metro-Vancouver area in British Columbia. Due to the COVID-19 pandemic, children participated remotely, and were individually tested on a Zoom call with an experimenter. Consent was obtained from the parent or legal guardian present at the time of the study, and experimenters received verbal assent from children. Participants were compensated with a $5 gift card for Amazon or Indigo.

Stimuli and Design. As the experiment was conducted using Zoom, a video teleconference platform, participants completed the study from home on their own computers. An experimenter ran a custom Psychopy (Peirce et al., 2019) script via Pavlovia, a website for hosting and running online experiments, and shared their screen to display visual stimuli to accompany the verbal interview questions.

Each participant received the same 5 practice trials to familiarize them with the mode of response, followed by 24 test trials in randomized order. The entire procedure lasted between 15 to 40 minutes, depending on the speed at which the child responded.

Within a trial, participants were asked two questions, and then were asked to make a forced-choice metacognitive judgment regarding which of their two answers was better. We included a range of question types (animal fact, animal ID, and numerical comparison; see Tables 10 and 11 in the Appendix for the list of practice and test questions) to examine explicit metacognitive judgments, and verbal disfluency, across contexts. Animal fact questions required children to retrieve fact-based knowledge from long-term memory, animal ID questions asked children to name an image of an animal, and numerical comparison questions showed children images of non-symbolic dot displays, with spatially separated collections of yellow dots on the left side of the screen and blue dots on the right and asked them which of the two sides had more dots, the blue or the yellow. We paired these questions so that—within a trial—the children answer two questions of the same type (e.g., two animal fact questions). We varied the difficulty of these paired questions in an effort to induce a range of certainty in their answers (e.g., within a trial, children are asked "What are baby cats called?" and "What are baby swans called?" and are then asked which of their two answers they felt most certain about). The difficulty of these questions was estimated before the study – we designed our questions so that children received the same number of pre-defined "easy", "mid", and "hard" questions (see Tables 10 and 11 in the Appendix for these questions by bin), and these pre-defined bins were validated post-hoc. While the order of the trials was randomized across participants, the question-pairs within trials was consistent across participants.

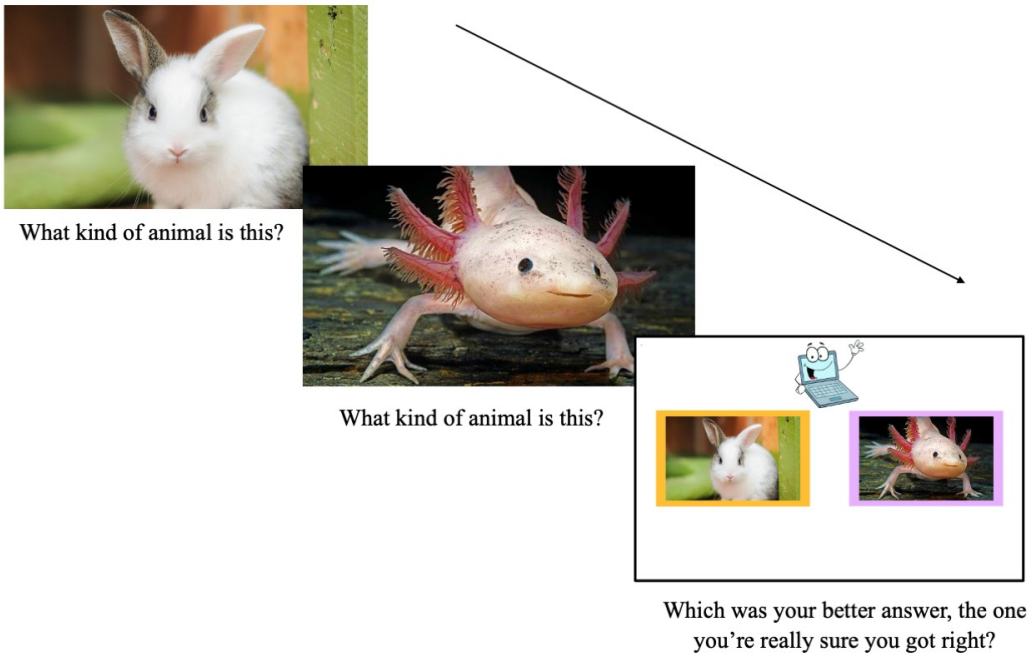
Procedures. At the time of the study, participants signed onto a private Zoom room. The experimenter reviewed the procedure with the child and their caregiver, before sharing their screen. Caregivers confirmed that the stimuli were properly displayed before beginning the experiment.

First, the experimenter explained the procedure to the child: they told the child that they would be answering a few different types of questions and emphasized that they child may not know the answer to all of them. The experimenter explained that the child should do their best to answer even when they are unsure. Then, they proceeded with the first practice trial. The experimenter displayed an image of a birthday cake and asked the child how old they are. Then, they displayed an image of a stranger, and asked the child how old the stranger is. After the child answered both questions, the experimenter introduced the procedure for the forced-choice metacognitive judgment. The images of the birthday cake and the stranger were displayed side by side. The birthday cake was surrounded by an orange outline, and the stranger by a purple outline. The experimenter explained that to "win" this game, the child needed to get a lot of questions right, and that to help the kid out, they will allow the child to answer two questions, and pick which of the two was their best answer, and they will choose this one for the computer to "check". Then, they asked the child "which of the two questions was their best answer, the one you're really sure you got right – the orange or the purple question?". The child indicated their response verbally by saying "purple" or "orange", and the experimenter pressed a key to record their response for analysis (see Figure 1 for the progression of trials). Then, the child answered 4 practice trials (see Table 10 in the Appendix for practice questions). All subsequent trials followed the same procedure, where the experimenter displayed an image associated with each question, then after the child answered both questions in the pair, displayed the pair of images

and asked the child to indicate which of the two was their better answer. These practice trials were designed to introduce the child to the various types of questions they will be answering at test.

Figure 1

Progression of the trial



Following the four practice trials, participants completed the 24 test trials in random order (see Table 11 in the Appendix for test questions), following the same procedure outlined above. Participants were permitted to skip any questions they did not know the answer to but were encouraged to provide their best guess. If the child was reluctant to answer, and did not provide a response within 15 seconds, the experimenter reminded them they could take a guess if they were unsure. If they did not answer after 30 seconds, the experimenter reminded them that they could say I don't know. If children skipped both knowledge questions (i.e., responded "I

don't know" to both questions), the experimenter omitted the forced-choice metacognitive judgment for that trial.

Coding. Audio was recorded during the testing session, using Zoom's audio recording feature. After testing, coders transcribed using the annotation software ELAN (Lausberg & Sloetjes, 2009). Coders transcribed the child's response using a custom coding template which allowed them to record timestamps as well categorize the child's answer into parts of speech of interest, including fillers, incomplete words, hedging phrases, and reinforcing phrases (see Table 1 for definitions and examples of category). This transcription allowed us to pull count and duration measures for each part of speech during analysis.

Table 1

Coding scheme for transcribing in ELAN

Part of Speech	Definition	Examples
Word	Any real speech that is part of the child's answer, not otherwise captured by our coding scheme	*"dog"; "pink"; "the yellow side"*
Hedge Phrases	Words or phrases which express uncertainty or hedging	*"I think"; "maybe"; "sort of"; "I guess"*
Reinforcing Phrases	Words or phrases which express certainty or reinforcing	*"I know"; for sure"; "definitely"*
Filler	Non-lexical sounds, filled pauses, or interjections	*"um"; "uh"; "eh"; "hmm"; "ah"*
Incomplete Word	Incomplete utterances – e.g, in the cases of stuttering or stammering	*"ra-"; "r-"*
Repeating Question	The child repeats the experimenter's question	*"Which side has more dots?"*
Non-answer	Phrases that are used to express a non-answer	*"I don't know"; "I have no guess"; "pass"*
Animal Sound	Used to tag sounds produced in answering "what sound does [animal] make?"	*"Roar"; "glub glub"; "grrr"*

Measures. Beyond the count and duration measures of each part of speech, we post-hoc calculate the mean duration of individual items for each part of speech within a trial (e.g., the mean length of individual fillers—e.g., ums and uhs—within a trial). We also standardize duration variables relative to the total duration of the child's answer, to see what proportion of their answer is devoted to fillers, for example. This standardization assuages concerns that longer responses simply allow for more randomly generated disfluency. See Table 2 for a definition of each dependent variable included in our analyses and descriptive statistics.

Table 2

Descriptive statistics for all measures. Means are calculated from participant-level averages across the experiment.

Measure	How we derive this measure	Mean [95 % CI]	Association with age
Speech Onset	Latency to begin answering (i.e., the silent period before which the child generates their first word)	1.67 [1.44, 1.90]	$r(58)= .21$, $p= .11$
Speech Offset	Time the child finished answering (i.e., response time).	4.24 [3.64, 4.84]	$r(58)= -.06$, $p= .66$
Filler Count	Count of individual filler terms (within an answer)	.26 [.20, .31]	$r(58)= -.25$, $p= .05$
Filler Duration	Cumulative duration of all fillers (within an answer)	.16 [.12, .20]	$r(58)= -.26$, $p= .05$
Mean Filler Duration	Filler duration/filler count	.57 [.51, .62]	$r(56)= -.14$, $p= .29$
Standardized Filler Duration	Filler duration/Speech offset	.03 [.02, .04]	$r(58)= -.29$, $p=.02$
Hedge Count	Count of individual hedge phrases (within an answer)	.10 [.07, .13]	$r(58)= .23$, $p= .08$
Hedge Duration	Cumulative duration of all hedge phrases (within an answer)	.08 [.05, .11]	$r(58)= .12$, $p= .38$
Mean Hedge Duration	Hedge count/Hedge duration	.82 [.64, 1.00]	$r(43)= -.33$, $p= .03$
Standardized Hedge Duration	Hedge duration/Speech Offset	.01 [.01, .02]	$r(58)= .12$, $p= .36$
Non-answer count	Count of non-answer phrases (within an answer)	.06 [.04, .08]	$r(58)= -.15$, $p= .24$
Accuracy	Whether their answer was correct. 0=incorrect, 1=correct.	.53 [.51, .56]	$r(58)= .52$, $p<.001$
Confidence Choice	Whether they chose this trial in the forced-choice metacognitive judgment. 0=not chosen, 1=chosen	.50	NA

Results

How accurate were children's answers?

To elicit a range of confidence responses, children must answer a set of questions that vary in difficulty, so that they answer some correctly, and others incorrectly. Before we begin to examine children's confidence judgments or their disfluency, therefore, we must look at the accuracy of their responses. Children, on average, answered 53.34% of questions correctly, 95% CI [50.96, 55.72]. Children's age correlated with their accuracy $r(58)= .52$, $p<.001$**.** Binning age by year, we see a linear increase in accuracy: 5-year-olds answered 46.18% (95% CI [40.90, 51.45]) of questions correctly, 6-year-olds answered 51.56% (95% CI [47.94, 55.17]) of questions correctly, 7-year-olds answered 57.06% of questions correctly (95% CI [51.65, 62.47]), and 8-year-olds answered 59.48% of questions correctly (95% CI [56.10, 63.85]).

Recall that we designed questions to vary in difficulty, so that children received an equal number of pre-defined "easy", "mid", and "hard" questions, and included a range of question types (numerical comparison, animal identification, and animal fact) (see Tables 10 and 11 in the Appendix for these questions by bin, and accuracy rates for specific questions). A repeated-measures ANOVA with accuracy as the dependent variable and two factors (Difficulty: Easy/Mid/Hard and Task: Numerical Comparison/Animal Identification/Animal Sound) revealed a significant main effect of Difficulty: participants provided more accurate answers for easier questions (easy: M=90.49%, 95% CI [88.16, 92.82]; mid: M=61.53%, 95% CI [57.14, 65.92]; hard: M=8.73%, 95% CI [6.50, 10.95]), $F(2,118)=854.35$, $p<.001$, $\eta_p^2=0.94$) and a main effect of Question Type (numerical comparison: M=62.41%, 95% CI [55.82, 69.00]; animal identification: M=53.67%, 95% CI [47.88, 59.45]; animal fact: M=44.68%, 95% CI [39.77, 49.58]), $F(2,118)=59.02$, $p<.001$, $\eta_p^2=0.50$. Furthermore, we observed an interaction between

Difficulty (easy, mid, hard) and Question Type (numerical comparison, animal fact, animal ID) $F(4,236)=80.00$ $p<.001$, $\eta_p^2=0.58$ (see Table 3). While participants provided overall more accurate answers for numerical comparison questions, they also provided the *least* accurate answers for hard questions in this category. This is likely due to differences in the structure of these questions: numerical comparison questions are binary choices, and are perceptual in nature, as the stimulus itself (the dot display) provides the child with possible answers. This binary choice makes these questions easier to answer, *except* for when neither option is correct: for hard numerical comparison questions, both sides of the display have the same number of dots.

Table 3

Accuracy rates by question type and difficulty bin.

Question Type	Difficulty Level		
	Easy	**Mid**	**Hard**
Numerical Comparison	95.56% [92.82, 98.29]	89.33% [84.44, 94.23]	2.33% [-0.90, 5.07]
Animal ID	97.00% [94.91, 99.09]	51.67% [45.34, 57.99]	12.33% [7.29, 17.38]
Animal Fact	78.92% [73.90, 83.93]	43.60% [39.64, 49.56]	11.51% [8.32, 14.71]

Do children's metacognitive judgments using the forced-choice paradigm track accuracy?

Next, we addressed whether the forced-choice judgment task is a valid assessment of children's metacognition. If children reliably report their confidence using this task, their responses should track the accuracy of their answer: wherein they choose the questions they are more likely to have answered correctly in the forced-choice paradigm.

A logistic regression with accuracy as the dependent variable and Trial Choice in the forced-choice metacognitive judgment as the explanatory revealed a significant effect of Trial Choice (z = 21.87, p < .001) with a odds ratio of 4.68 (95% CI [4.08, 5.38]), indicating that participants were about 4.5 times more likely to be accurate on the trials they chose (M=72.33%, 95% CI [68.77, 75.90]) compared to ones they did not (M=35.08%, 95% CI [32.25, 37.92]). This was true for all question types: numerical comparison (z=11.12, p<.001, OR=4.87, 95% CI [3.70, 6.47]; $M_{Chosen\ trials}$=81.85%, 95% $CI_{Chosen\ trials}$ [77.43, 86.26]; $M_{Not\ chosen\ trials}$=47.46%, ,95% $CI_{Chosen\ trials}$ [44.16, 50.77]), animal identification (z=12.79, p<.001, OR=6.49, 95% CI [4.89,8.67]; $M_{Chosen\ trials}$=75.67%, 95% $CI_{Chosen\ trials}$ [70.74, 80.60]; $M_{Not\ chosen\ trials}$=33.31%, 95% $CI_{Chosen\ trials}$ [28.06, 38.59]), and animal fact (z=13.11, p<.001, OR=3.83, 95% CI [3.13, 4.69]; $M_{Chosen\ trials}$=62.52%, 95% $CI_{Chosen\ trials}$ [57.73, 67.32]; $M_{Not\ chosen\ trials}$=30.63%, ,95% $CI_{Chosen\ trials}$ [26.46, 34.81]).

To determine any age-related effects – as metacognitive judgments may improve and become more accurate with development – we looked to individual differences in metacognitive sensitivity. For each participant, we calculated the percent of trials where they chose the objectively easier trial in the pair (as in Baer & Odic, 2019). Metacognitive sensitivity and age were moderately correlated (r(58)=0.42, p<.001) suggesting metacognitive sensitivity increased with age, consistent with prior work (Baer & Odic, 2019).

Do children produce verbal disfluency?

Children produced a range of disfluency across the experiment. Two children produced no non-lexical disfluency at all (i.e., used a hedge phrase but no fillers), but most children did use disfluency when answering our questions (Fillers: range= 0-70, M=16.58 , 95% CI [12.92,

20.24; Hedges: range= 0-24, *M*=6.57*,* 95% CI [4.61,8.52]). See Table 12 in the Appendix for individual differences in the amount of disfluency. The number of filler disfluencies across the experiment and age were moderately negatively correlated ($r(58)= -.25$, $p= .05$), suggesting that older children produced less filler disfluency than younger children. Children also produced various token forms of disfluencies and hedges. Most commonly, children used fillers "um", "uh", and "mmm" and hedges "I think" and "like". See Tables 4 and 5 for the count of individual disfluency and hedge tokens, across children, and across the experiment.

Table 4

Count of individual filler tokens, across participants and the duration of the experiment.

Filler	N
uh	390
um	249
mmm	177
hmm	45
ooh	13
eh	11
ah	11
sigh	8
ugh	4
huh	3
em	2
ay	2
er	2
yep	1

Table 5

Count of individual hedge tokens, across participants and the duration of the experiment.

Hedge	N
I think	113
like	36
maybe	21
kind of	15
probably	12
ish	7
could be	2
or something	2
Sort of	1
I guess	1
it's either	1

Next, we assessed whether disfluency rates differed across question types (numerical comparison, animal identification, animal fact). If children produced differing amounts of disfluency across question types, then this would warrant separate analysis for each question type. However, a one-way ANOVA with disfluency count as the dependent variable and Question Type as a factor revealed no significant effect of Question Type, $F(2,118)=2.68$, $p = .07$, $\eta_p^2=0.04$.

Do children's verbal disfluencies track their accuracy?

Subsequently, we determined how children's verbal disfluencies related to the accuracy of their response. We performed a series of logistic regressions to examine the effects of all our measures of disfluency in predicting accuracy (see Table 2 for definitions of measures). We observed a similar pattern across all measures—excepting mean hedge duration: children

produced more fillers and hedges in answering questions incorrectly. Furthermore, they took longer to provide incorrect answers, evidenced by speech onset (see Table 6).

Table 6

Disfluency rates by accuracy.

Measure	Correct Answers	Incorrect Answers	Stats
Filler Count	.17 [.13, .21]	.36 [.28, .44]	z=-9.56, p<.001, OR=0.54, 95% CI [.47, .61]
Filler Duration	.09 [.07, .12]	.23 [.17, .29]	z=-10.30, p<.001, OR=.35, 95% CI [.29, .43]
Mean Filler Duration	.57 [.50, .65]	.58 [.52, .64]	z=-3.92, p<.001, OR=.39, 95% CI [.24, .62]
Standardized Filler Duration	.02 [.02, .03]	.04 [.03, .05]	z=-7.32, p<.001, OR=.03, 95% CI [.01, .06]
Hedge Count	.05 [.03, .07]	.17 [.12, .22]	z=-7.77, p<.001, OR=.44, 95% CI [.36, .54]
Hedge Duration	.04 [.02, .06]	.13 [.07, .18]	z=-6.30, p<.001, OR=.48, 95% CI [.38, .60]
Mean Hedge Duration	.81 [.55, 1.08]	.81 [.63, .99]	z=-.90, p=.37, OR=.85, 95% CI [.56,1.25]
Standardized Hedge Duration	.01 [.00, .01]	.02 [.01, .03]	z=-5.48, p<.001, OR=.03, 95% CI [.01, .10]
Speech Onset	1.19 [1.03, 1.36]	2.29 [1.92, 2.65]	z=-12.90, p<.001, OR=.76, 95% CI [.73, .79]

Next, to assess the cumulative effects of various measures, we performed a logistic regression that includes several measures of disfluency. Given count, duration, mean duration, and standardized duration measures are not independent, we selected standardized duration as our primary variable of interest. A series of hierarchical logistic regressions revealed that a cumulative disfluency model (including standardized filler duration, standardized hedge

duration, and speech onset as explanatory variables) best fit our data, and jointly predicted accuracy significantly better than any of them did independently (see Table 7).

Table 7

Hierarchical logistic regression comparison. A cumulative model of disfluency best fit our data.

	Model 1	**Model 2**	**Model 3**
Standardized Disfluency Duration	z=-7.32, p<.001, OR=.03, 95% CI [.01, .06]	z=-7.16, p<.001, OR=.03, 95% CI [.01, .07]	z=-8.12, p<.001, OR=.02, 95% CI [.01, .05]
Standardized Hedge Duration		z=-5.31, p<.001, OR=.04, 95% CI [.01, .11]	z=-5.57, p<.001, OR=.02, 95% CI [.01, .10]
Speech Onset			z=-13.37, p<.001, OR=.75, 95% CI [.72, .78]
ΔAIC	0	36.13	246.20

Do children's verbal disfluencies track their metacognitive awareness?

To assess how disfluency tracks explicit metacognitive judgments, we performed a series of Welch's t-tests (given unequal variances) to examine all our measures of disfluency across chosen and not-chosen trials in the forced-choice metacognitive judgment. Again, we observed a similar pattern across all measures – children produced more fillers and hedges, and took longer to begin answering, during the trials they did not choose in the forced-choice comparison (see Table 8).

Table 8

Disfluency rates by explicit metacognitive judgment (chosen vs. not chosen trials in the forced-choice comparison)

Measure	Chosen Trials	Not-chosen Trials	Stats
Filler Count	.19 [.14, .24]	.32 [.25, .39]	$t(3614.2)=6.89$, $p<.001$, $d=0.22$
Filler Duration	.11 [.08, .14]	.21 [.16, .26]	$t(3589.2)=7.56$, $p<.001$, $d=0.24$
Mean Filler Duration	.55 [.49, .60]	.66 [.55, .66]	$t(631.41)=3.40$, $p<.001$, $d=0.25$
Standardized Filler Duration	.02 [.01, .03]	.04 [.03, .05]	$t(3585)=6.51$, $p<.001$, $d=0.21$
Hedge Count	.06 [.04, .08]	.15 [.10, .19]	$t(3291.5)=6.78$, $p<.001$, $d=0.22$
Hedge Duration	.04 [.02, .06]	.12 [.07, .17]	$t(3002.8)=6.69$, $p<.001$, $d=0.22$
Mean Hedge Duration	.73 [.51, .94]	.85 [.66, 1.04]	$t(173.73)=2.08$, $p=.04$, $d=0.25$
Standardized Hedge Duration	.01 [.00, .01]	.02 [.01, .03]	$t(3281)=5.14$, $p<.001$, $d=0.17$
Speech Onset	1.40 [1.22, 1.62]	1.94 [1.67, 2.24]	$t(3527.6)=7.38$, $p<.001$, $d=0.24$

What is the best measure of children's metacognition?

Finally, we assessed whether disfluency predicts accuracy over and above their explicit confidence judgments. Recall that we observed children's explicit metacognitive judgments predicted the accuracy of their responses, as did their disfluency (see Table 6). First, we compared which of these two models best fit our data, using AIC model comparison $AIC_{\text{disfluency model}} - AIC_{\text{metacognitive judgment model}}=151.57$, thus suggesting that disfluency predicts accuracy better than children's metacognitive judgments (disfluency model: $z_{\text{standardized filler duration}}=-8.12$, $p<.001$, OR=.02, 95% CI [.01, .05]; $z_{\text{standardized hedge duration}}=-5.57$, $p<.001$, OR=.02, 95% CI [.01, .10];

$z_{\text{speechonset}}$ =-13.37, p<.001, OR=.75, 95% CI [.72, .78]; metacognitive judgment model: $z_{\text{trial chosen}}$=21.87, p < .001, OR=4.68, 95% CI [4.08, 5.38]).

Next, we performed a logistic regression to assess the cumulative effects of disfluency and explicit metacognitive judgments in predicting accuracy, entering trial chosen, standardized filler duration, standardized hedge duration, and speech onset as predictor variables ($z_{\text{trial chosen}}$=19.68, p < .001, OR=4.22, 95% CI [3.66, 4.87]; $z_{\text{standardized filler duration}}$=-6.55, p<.001, OR=.03, 95% CI [.01, .09]; $z_{\text{standardized hedge duration}}$=-4.67, p<.001, OR=.06, 95% CI [.02, .19]; $z_{\text{speechonset}}$ =-11.99, p<.001, OR=.76, 95% CI [.73, .80]). This final model revealed a significantly better fit, suggesting that disfluency and metacognitive judgments jointly predict accuracy better than either measure alone ($AIC_{\text{metacognitive judgment model}}$ - $AIC_{\text{metacognitive judgment and disfluency combined model}}$=255.77; $AIC_{\text{disfluency model}}$ - $AIC_{\text{metacognitive judgment and disfluency combined model}}$=407.34).

Exploratory Analyses

Disfluency rates across answers and non-answers.

To assess whether disfluency rates differ across answered and unanswered questions (i.e., questions children responded "I don't know" to), we performed a series of t-tests across answered and unanswered questions. We observed a significant effect of answer type: kids produced more fillers during questions they ultimately provided no answer for (see filler count and duration, table 9). The duration of individual fillers, however, was longer for answered questions compared to unanswered ones, as was the cumulative and standardized duration of hedge phrases (see table 9).

Table 9

Disfluency rates by answer type (answered vs. unanswered trials)

Measure	Answered Trials	Unanswered Trials	Stats
Filler Count	.24 [.19, .30]	.55 [.34, .76]	$t(210.71)=-3.09$, $p=.002$, $d=0.34$
Filler Duration	.16 [.12, .19]	.29 [.17, .42]	$t(211.95)=-2.21$, $p=.03$, $d=0.22$
Mean Filler Duration	.59 [.53, .64]	.51 [.40, .62]	$t(68.87)=2.65$, $p=.01$, $d=0.34$
Standardized Filler Duration	.03 [.02, .04]	.03 [.02, .03]	$t(221.37)=-.47$, $p=.64$, $d=0.04$
Hedge Count	.10 [.07, .13]	.12 [.00, .24]	$t(229.41)=1.28$, $p=.20$, $d=0.08$
Hedge Duration	.08 [.05, .11]	.07 [.01, .13]	$t(268.89)=2.34$ $p=.02$, $d=0.11$
Mean Hedge Duration	.82 [.64, 1.00]	.70 [.28, 1.12]	$t(9.65)=1.43$, $p=.18$, $d=0.28$
Standardized Hedge Duration	.01 [.01, .02]	.00 [.00, .01]	$t(897.41)=7.43$, $p<.001$, $d=0.17$
Speech Onset	1.66 [1.43, 1.89]	1.97 [1.37, 2.57]	$t(212.78)=-1.66$, $p=.10$, $d=0.17$

Individual differences in non-answer rates.

Although we encouraged participants to guess when they were unsure, children varied in their willingness to guess. Some children, when they did not know the answer to a question, would take a guess, or describe adjacent concepts (e.g., one child, when identifying an axolotl, responded "I think it's a chamele-, a funny chameleon with, uh, spikes on its sides"). Other children would answer "pass" or "I don't know". Thus, participants differed in their rate of non-answers (see Table 12 in the Appendix for individual differences in the rate of non-answer phrases. Note that, for these analyses, we collapse non-answer phrases into a binary score within

individual trials, i.e., “I don’t know, pass”, although two non-answer *phrases,* is one *non-answer*). Given the differences in disfluency rates and non-answers, we might find that individual differences in participants non-answer rate predicts their amount of disfluency across the experiment. A series of regressions revealed that individual differences in non-answer rate predicted individual differences in speech onset: children who answered fewer questions took longer to begin answering, R^2=.07, F(1,51)=3.96, β=1.47, p=.05. No effects of non-answer rate were observed for our other measures of disfluency (filler count: R^2=.01, F(1,58)=.87, β=.27, p=.36; filler duration: R^2=.02, F(1,58)=1.23, β=.18, p=.27; mean filler duration: R^2=.04, F(1,56)=2.17, β=.60, p=.15; standardized filler duration: R^2=.009, F(1,58)=.50, β=.03, p=.48; hedge count: R^2=.001, F(1,58)=.04, β=.10, p=.84; hedge duration: R^2=.001, F(1,58)=.07, β=.08, p=.79; mean hedge duration: R^2=.001, F(1,43)=.03, β=.83, p=.86; standardized hedge duration: R^2=.001, F(1,58)=.06, β=.01, p=.81).

Gender effects.

A series of t-tests on all our measures of disfluency revealed effects of gender for filler count and duration, and hedge count. Girls, on average, produced more and longer fillers than boys (filler count; girls: M=0.29, 95% CI [.20, .37], boys: M=0.21, 95% CI [.15, .28]; t(3431.9)= -3.52, p <.001; filler duration; girls: M=0.17, 95% CI [.11, .23], boys: M=0.14, 95% CI [.09, .19]; t(3691.4)= -2.43, p=.02; and more hedges (hedge count; girls: M=0.12, 95% CI [.07, .17], boys: M=0.08, 95% CI [.04, .12] t(3553.4)= -2.61, p=.000; hedge duration; girls: M=0.09, 95% CI [.03, .15], boys: M=0.06, 95% CI [.02, .10] t(3353.6)= -2.43, p=.02)

Um/uh distinction.

Recall that Smith & Clark (1993) found a distinction between “um” and “uh”: participants used “uh” to signal short delays, and “um” for longer ones. Thus, we examine the length of the pause following “ums” and “uhs”. A t-test revealed an insignificant trend toward longer pauses following “ums” (M=1.50, 95% CI [.83,2.17]) than “uhs” (M=0.95, 95% CI [.47, 1.44]), $t(237.75)=1.90$, $p = .06$, $d=0.22$.

Discussion

This study sought to examine the relationship between produced verbal disfluency and confidence in young children. We asked children questions about animals and number and looked at how their disfluency relates to the objective accuracy of their response and their subjective explicit ratings of confidence. We found that children produced more disfluency when they answered questions incorrectly, and when they indicated feeling less confident. I will discuss these results with respect to the four primary research questions from the introduction.

(1) Do English-speaking preschool-aged children produce verbal disfluencies and, if so, what kinds?

Excepting two participants, all children in our study produced non-verbal disfluency in answering our questions. There were individual differences in the amount of disfluency children produced, however: kids produced between 0 and 70 fillers, and between 0 and 24 hedge phrases across the experiment (see Table 12 in the Appendix). These individual differences were partially predicted by gender: girls, on average, produced more fillers and hedges than boys, as well as age. We observed a decrease in filler disfluencies with age: older children produced less fillers than younger children. It is possible that this is because older children have more practice using language, and therefore are simply more fluent speakers. Alternatively, it is possible that this age-related decrease in disfluency is a consequence of the age-related increase in accuracy. Given that older children provided more accurate answers, and that the rate of disfluency tracks the accuracy of children's responses, then it follows that older children who produce more accurate answers would also produce less verbal disfluency.

However, while fillers decreased with age, hedge phrases increased, suggesting older children may more often explicitly mark uncertainty in their answers. This is similar to Catalan-speaking children, who used implicit prosodic and gestural cues before lexical markers (Hübscher et al., 2019). This suggests that – although we observed both fillers and hedges to predict the accuracy of children's answers, as found in adults (Smith & Clark, 1993) – hedges may not always be developmentally appropriate. For the younger children in our sample (5-year-olds) non-lexical disfluencies may be a more ecologically valid measure than hedges. In other words, looking at hedges alone may be an insufficient measure of uncertainty for children under 5.

Children produced a range of token disfluency terms, too, though most commonly they used non-lexical fillers "um" and "uh" (see Table 4), much like English-speaking adults (Smith & Clark, 1993). While children also used hedges, although at an overall lower rate than they used fillers, as found in Smith & Clark (1993), they used different token hedge phrases than adults did. Adults often hedged their answers with "I guess", but this was rather uncommon in children's speech. Children most frequently used "I think" to hedge their answers (see Table 5). In a previous case study of the acquisition of mental verbs, one child used "think" to express mental state functions prior to "guess" (Shatz et al., 1983). It is possible that "I think" is therefore acquired prior to other hedge phrases, however, given that this prior study relied on a longitudinal observation of only one child, future work is needed to determine the generalizability of these results.

To further evaluate whether children use disfluencies like adults, we assessed whether they differentiate between individual non-lexical filler terms "um" and "uh". Recall that Smith and Clark (1993) observed that participants reliably used "um" to signal longer delays an "uh" to

signal short pauses. We found that children did not differentially use “um” and “uh” as reliably: we observed an insignificant trend toward longer pauses following “ums” compared to “uhs” (see exploratory analyses in the Appendix). Similarly, Hudson Kam and Edwards (2008) found that younger children than those in the present study—3- and 4-year-olds—did not use “um” and “uh” to signal different degrees of disruption. The authors propose that this semantic differentiation may occur with further learning and development, wherein children learn the unique meaning of “um” and “uh” with experience (Hudson Kam & Edwards, 2008). Indeed, it is possible that our trend toward longer pauses following ums reflects age-related changes: that the older children in our sample—5-to-8-year-olds—have begun to differentiate between “ums” and “uhs”, but more work is necessary, with a larger age range, to determine this developmental trajectory.

Alternatively, it is possible that this is not a semantic distinction that needs to be learned. While some argue that this difference in the length of pauses necessitates that fillers function like any other word: that we selectively use “um” to signal a greater challenge compared to “uh”, it is possible that this association is a natural consequence of speech production. An “um” can be considered simply an “uh” with an “m” attached to the end. It is possible that a longer pause provides us greater time to terminate an “uh” with an “m”. Indeed, while the length of pauses differs following “ums” and “uhs” (Smith & Clark, 1993), Brennan & Williams (1995) did not find that these contrast in speech perception: the presence of an “um” relative to an “uh”, controlling for the length of pauses, did not influence participants’ ratings of another’s knowing. This suggests that we do not explicitly differentiate between the semantic meaning of “um” and “uh”, evidence against accounts which view fillers as morphemes (e.g., Clark & Fox Tree, 2002).

More studies on young children's perception of disfluency are needed to determine whether children distinguish between the meaning of "um" and "uh".

(2) **Do disfluencies that children produce relate to the accuracy of their responses? In other words, could disfluencies be an *implicit* measure of confidence states?**

We found that children took longer to begin answering and produced more disfluencies when providing objectively incorrect answers compared to correct ones, aligned with results among Dutch- (Krahmer & Swerts, 2005; Visser et al., 2014) and Catalan-speaking (Hübscher et al., 2019) children. This was true across all but one of our measures of disfluency—mean hedge duration (although, note that other hedge variables—hedge count, hedge duration, and standardized hedge duration—did track the accuracy of children's response). It is possible that lexical disfluencies are less often manipulated by phonological devices like sound stretch compared to non-lexical items. Accordingly, the duration of individual hedge tokens may be overall less variable than individual filler tokens, for example.

For all other indices of disfluency, however—filler count, filler duration, mean filler duration, standardized filler duration, hedge count, hedge duration, standardized hedge duration, and speech onset—means were higher for inaccurate answers compared to accurate ones. Notably, this suggests that children not only took longer to begin answering, and produced more fillers and hedge phrases, but also that they produced longer *individual* fillers when they felt less confident.

Thus, it appears that various *types* of disfluencies can all be considered markers of accuracy: it does not appear that either hedges, fillers, or pauses should be considered better or worse indices. These various disfluencies, however, should not be considered simply

confounding measures: while they all track accuracy similarly, a logistic regression including several categories of disfluency (standardized filler duration, standardized hedge duration, and speech onset) predicted accuracy significantly better than any measure did independently, and confirmed that these variables were not collinear.

In sum, it does appear that disfluencies could be an *implicit* measure of confidence states in children. Furthermore, that we observed this effect across all our measures of disfluency suggests that future work should consider *all* categories of disfluencies—fillers, hedges, and pauses—as an index of the probability of success.

(3) **Do disfluencies that children produce relate to the explicit confidence in their responses?**

If disfluency can be considered an *implicit* measure of confidence, how do they relate to children's *explicit* ratings of confidence? This question is particularly interesting in younger children, as previous work has suggested dissociations between disfluency and explicit metacognitive judgments (Hübscher et al., 2019; Visser et al., 2014), but given these tasks relied on Likert scale tasks which allow children to respond "very confident" on every trial, these paradigms may have confounded metacognitive sensitivity for bias.

Instead, we used a forced-choice task that removes children's ability to repond overconfidently. We found that children could reliably report their metacognition using this task: they were significantly more accurate on the trials they chose in the forced-choice metacognitive judgment relative to the trials they did not choose, aligned with prior work using this paradigm (Baer et al., 2018, 2021; Baer & Odic, 2019, 2020; Butterfield et al., 1988). This was true for all

of our different question types (numerical comparison, animal identification, animal fact), suggesting this method is appropriate for assessing confidence across domains.

Using a task that controls for bias, we found that, across all our measures of disfluency—filler count, filler duration, mean filler duration, standardized filler duration, hedge count, hedge duration, mean hedge duration, mean standardized hedge duration, and speech onset—disfluency tracked their forced-choice decisions. Children produced more disfluency on trials that they did not choose in the forced-choice metacognitive judgment, relative to the trials they chose.

While Krahmer and Swerts (2005) observed a similar effect in Dutch-speaking children, they found that children did *not* reliably use fillers to mark their uncertainty: there were no significant differences in children's FOK ratings in the presence/absence of fillers. The children in our study, however, *did* use fillers to mark their uncertainty, much like adults (Smith & Clark, 1993). This discrepancy may reflect cross-linguistic differences in children's use of disfluency, or methodological limitations of the Likert scale metacognitive judgment task of Krahmer and Swerts (2005).

(4) Are disfluencies a *better* predictor of accuracy than explicit confidence judgements, even when controlling for bias?

Both children's explicit metacognitive judgments and their disfluency tracked the accuracy of their response, but our series of models showed that disfluency *better* predicted accuracy than their explicit judgments of confidence. Furthermore, disfluency accounted for additional variance over and above confidence judgments alone. In other words, children's accuracy was best predicted by a combination of confidence judgments *and* verbal disfluency. This suggests that, even when controlling for bias, and using a metacognitive judgment task that

facilitates reports for younger children, implicit measures may serve as a better window into children's confidence than explicit ratings. Thus, future work should consider assessing disfluency, rather than explicit judgments, as a measure of children's metacognition.

Other implicit measures of metacognition, such as opting out of difficult trials (Balcomb & Gerken, 2008; Bernard et al., 2015; Ghetti et al., 2013; Lyons & Ghetti, 2013), facial expression (van Amelsvoort et al., 2013), or eye-tracking (Paulus et al., 2013), have been increasingly favored over explicit measures for children of this age, as these implicit measures are less susceptible to response biases, like children's overconfidence or "wishful thinking" (Schneider, 1998). Disfluency may provide an even better implicit measure, however, as it can be combined with nearly any method in psychology that relies on verbal report, and is more accessible than eye-tracking or facial expression methods, which require complex technologies (i.e., eye-trackers or automated facial expression recognition software) or behavioral coding of facial expression, which is subjective and open to individual interpretation. Given advances in machine-learning speech recognition and coding technology, there is a high chance that future work could code disfluency automatically through these models.

Furthermore, this line of work can clarify the directionality between disfluency and explicit metacognition. Some views hold that metacognition is the perception of (dis)fluency: that we determine our level of confidence, explicitly, by how easily the information is retrieved from memory—or the perception of "cognitive fluency" (Alter & Oppenheimer, 2009; Koriat, 1993). Verbal disfluencies may play a similar role in metacognition: we may, for example, monitor our speech for the presence of "ums" and "uhs" to explicitly determine our level of confidence. The observed association between children's disfluency and their metacognitive judgments may be interpreted as support for this theory across development.

However, given prior work showing that children's disfluency can dissociate from their explicit metacognitive judgments (Hübscher et al., 2019; Visser et al., 2014), there is reason to suspect that one is not directly derived from the other. Explicit metacognitive judgments may be greater influenced by external task demands, like children's "wishful thinking" (Schneider, 1998), compared to verbal disfluency. Indeed, in one study, 11-year-old children made more high confident metacognitive judgments in competitive compared to collaborative contexts, suggesting they attempt to signal to others that they are well-matched competitors (Visser et al., 2014). Their verbal disfluency, in contrast, was not influenced by social context, and thus this implicit measure was less manipulable by social and task demands. If explicit judgments, but not verbal disfluency, are influenced by social context, then explicit metacognition is not *solely* determined by perceived disfluency. Future work is needed to examine the effects of social task demands on both disfluency and explicit judgments in younger children, determining, for example, the influence of experimenter presence on children's "wishful thinking" (Schneider, 1998). Furthermore, additional studies should examine the relationship between verbal disfluency and *prospective* confidence ratings: where children provide metacognitive judgments prior to answering the question (e.g., asking "which question do you want to answer" instead of "which question was your better answer", Baer et al., 2021). If explicit metacognition is perceived (dis)fluency, then it follows that children must answer the question *before* explicitly judging their confidence, as in the present study. Under this theory, it is unclear whether we would observe a similar pattern using a retrospective metacognitive judgment task.

In conclusion, we found that children produce disfluency like adults, wherein the presence of fillers, hedges, and pauses, correlated with the accuracy of their answers, and related to their explicit ratings of metacognition. Furthermore, children's verbal disfluencies predicted

the accuracy of their answers over and above their confidence judgments alone, suggesting that future work should consider disfluency as an implicit measure of confidence. Future studies should consider examining how verbal disfluency relates to not only explicit ratings of confidence, but also other implicit behavioral markers, like posture, facial expression, and eye-gaze. It would be interesting to assess whether these implicit cues similarly predict accuracy over and above explicit judgments. Together, these findings will further our understanding children's developing metacognitive reasoning.

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the Tribes of Fluency to Form a Metacognitive Nation. *Personality and Social Psychology Review*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Amiridze, N., Davis, B. H., & Maclagan, M. (2010). *Fillers, Pauses and Placeholders*. John Benjamins Publishing.

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies Signal Theee, Um, New Information. *Journal of Psycholinguistic Research*, *32*(1), 25–36. https://doi.org/10.1023/A:1021980931292

Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 914–930. https://doi.org/10.1037/0278-7393.33.5.914

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The Old and Thee, uh, New: Disfluency and Reference Resolution. *Psychological Science*, *15*(9), 578–582. https://doi.org/10.1111/j.0956-7976.2004.00723.x

Baer, C., Gill, I. K., & Odic, D. (2018). A Domain-General Sense of Confidence in Children. *Open Mind*, *2*(2), 86–96. https://doi.org/10.1162/opmi_a_00020

Baer, C., Malik, P., & Odic, D. (2021). Are children's judgments of another's accuracy linked to their metacognitive confidence judgments? *Metacognition and Learning*, *16*(2), 485–516. https://doi.org/10.1007/s11409-021-09263-x

Baer, C., & Odic, D. (2019). Certainty in numerical judgments develops independently of the approximate number system. *Cognitive Development*, *52*, 100817. https://doi.org/10.1016/j.cogdev.2019.100817

Baer, C., & Odic, D. (2020). Children flexibly compare their confidence within and across perceptual domains. *Developmental Psychology*, *56*(11), 2095. https://doi.org/10.1037/dev0001100

Bailey, K. G. D., & Ferreira, F. (2003). Disfluencies affect the parsing of garden-path sentences. *Journal of Memory and Language*, *49*(2), 183–200. https://doi.org/10.1016/S0749-596X(03)00027-5

Balcomb, F. K., & Gerken, L. (2008). Three-year-old children can access their own memory to guide responses on a visual matching task. *Developmental Science*, *11*(5), 750–760. https://doi.org/10.1111/j.1467-7687.2008.00725.x

Barr, D. (2001). Trouble in mind: Paralinguistic indices of effort and uncertainty in communication. *Undefined*. https://www.semanticscholar.org/paper/Trouble-in-mind%3A-paralinguistic-indices-of-effort-Barr/02ef0d8159d445fa2e7b346bc0145f2a59b1894a

Barr, D. J. (2003). Paralinguistic correlates of conceptual structure. *Psychonomic Bulletin & Review*, *10*(2), 462–467. https://doi.org/10.3758/BF03196507

Barr, D. J., & Seyfeddinipur, M. (2010). The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes*, *25*(4), 441–455. https://doi.org/10.1080/01690960903047122

Bayard, N. S., van Loon, M. H., Steiner, M., & Roebers, C. M. (2021). Developmental Improvements and Persisting Difficulties in Children's Metacognitive Monitoring and

Control Skills: Cross-Sectional and Longitudinal Perspectives. *Child Development*, *92*(3), 1118–1136. https://doi.org/10.1111/cdev.13486

Beňuš, Š., Gravano, A., & Hirschberg, J. (2011). Pragmatic aspects of temporal accommodation in turn-taking. *Journal of Pragmatics*, *43*(12), 3001–3027. https://doi.org/10.1016/j.pragma.2011.05.011

Bernard, S., Proust, J., & Clément, F. (2015). Procedural Metacognition and False Belief Understanding in 3- to 5-Year-Old Children. *PLOS ONE*, *10*(10), e0141321. https://doi.org/10.1371/journal.pone.0141321

Birch, S. A. J., Akmal, N., & Frampton, K. L. (2010). Two-year-olds are vigilant of others' non-verbal cues to credibility. *Developmental Science*, *13*(2), 363–369. https://doi.org/10.1111/j.1467-7687.2009.00906.x

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech*, *44*(2), 123–147. https://doi.org/10.1177/00238309010440020101

Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, *34*(3), 383–398.

Butterfield, E. C., Nelson, T. O., & Peck, V. (1988). Developmental aspects of the feeling of knowing. *Developmental Psychology*, *24*(5), 654–663. https://doi.org/10.1037/0012-1649.24.5.654

Chomsky, N. (1965). *Aspects of the theory of syntax*.

Clark, H. H. (2002). Speaking in time. *Speech Communication*, *36*(1), 5–13. https://doi.org/10.1016/S0167-6393(01)00022-X

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association. https://doi.org/10.1037/10096-006

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Cook, M., Smith, J., & Lalljee, M. G. (1974). Filled Pauses and Syntactic Complexity. *Language and Speech*, *17*(1), 11–16. https://doi.org/10.1177/002383097401700102

Corley, M., & Hartsuiker, R. J. (2011). Why Um Helps Auditory Word Recognition: The Temporal Delay Hypothesis. *PLOS ONE*, *6*(5), e19792. https://doi.org/10.1371/journal.pone.0019792

Crible, L. (2018). *Discourse Markers and (Dis)fluency: Forms and Functions Across Languages and Registers* (Issue Vol. 286). John Benjamins Publishing Company. https://ezproxy.cul.columbia.edu/login?qurl=https%3a%2f%2fsearch.ebscohost.com%2flogin.aspx%3fdirect%3dtrue%26AuthType%3dip%26db%3de025xna%26AN%3d1708701%26site%3dehost-live%26scope%3dsite

Crible, L., Degand, L., & Gilquin, G. (2017). The clustering of discourse markers and filled pauses: A corpus-based French-English study of (dis)fluency. *Languages in Contrast*, *17*(1), 69–95. https://doi.org/10.1075/lic.17.1.04cri

DeJoy, D. A., & Gregory, H. H. (1985). The relationship between age and frequency of disfluency in preschool children. *Journal of Fluency Disorders*, *10*(2), 107–122. https://doi.org/10.1016/0094-730X(85)90019-1

Destan, N., & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning*, *10*(3), 347–374. https://doi.org/10.1007/s11409-014-9133-z

Dizik, A. (2016, June 16). *The secret to stopping your 'ummms.'* BBC Worklife. https://www.bbc.com/worklife/article/20160615-the-secret-to-stopping-your-ummms

Engelhardt, P. E., Corley, M., Nigg, J. T., & Ferreira, F. (2010). The role of inhibition in the production of disfluencies. *Memory & Cognition*, *38*(5), 617–628. https://doi.org/10.3758/MC.38.5.617

Ferreira, F., & Bailey, K. G. D. (2004). Disfluencies and human language comprehension. *Trends in Cognitive Sciences*, *8*(5), 231–237. https://doi.org/10.1016/j.tics.2004.03.011

Finlayson, I. R., & Corley, M. (2012). Disfluency in dialogue: An intentional signal from the speaker? *Psychonomic Bulletin & Review*, *19*(5), 921–928. https://doi.org/10.3758/s13423-012-0279-x

Finn, B., & Metcalfe, J. (2014). Overconfidence in children's multi-trial judgments of learning. *Learning and Instruction*, *32*, 1–9. https://doi.org/10.1016/j.learninstruc.2014.01.001

Fox, B. A., Maschler, Y., & Uhmann, S. (2010). A cross-linguistic study of self-repair: Evidence from English, German, and Hebrew. *Journal of Pragmatics*, *42*(9), 2487–2505. https://doi.org/10.1016/j.pragma.2010.02.006

Fox Tree, J. E. (1995). The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. *Journal of Memory and Language*, *34*(6), 709–738. https://doi.org/10.1006/jmla.1995.1032

Fox Tree, J. E. (2006). Placing like in telling stories. *Discourse Studies*, *8*(6), 723–743. https://doi.org/10.1177/1461445606069287

Fox Tree, J. E., & Schrock, J. C. (1999). Discourse Markers in Spontaneous Speech: Oh What a Difference an Oh Makes. *Journal of Memory and Language*, *40*(2), 280–295. https://doi.org/10.1006/jmla.1998.2613

Ghetti, S., Hembacher, E., & Coughlin, C. A. (2013). Feeling Uncertain and Acting on It During the Preschool Years: A Metacognitive Approach. *Child Development Perspectives*, *7*(3), 160–165. https://doi.org/10.1111/cdep.12035

Goupil, L., & Kouider, S. (2019). Developing a Reflective Mind: From Core Metacognition to Explicit Self-Reflection. *Current Directions in Psychological Science*, *28*(4), 403–408. https://doi.org/10.1177/0963721419848672

Grice, P. (1989). *Studies in the Way of Words*. Harvard University Press.

Hayashi, M., & Yoon, K.-E. (2006). A cross-linguistic exploration of demonstratives in interaction: With particular reference to the context of word-formulation trouble. *Studies in Language. International Journal Sponsored by the Foundation "Foundations of Language," 30*(3), 485–540. https://doi.org/10.1075/sl.30.3.02hay

Hinton, L., Moonwomon, B., Bremner, S., Luthin, H., Van Clay, M., Lerner, J., & Corcoran, H. (1987). It's Not Just the Valley Girls: A Study of California English. *Annual Meeting of the Berkeley Linguistics Society*, *13*, 117. https://doi.org/10.3765/bls.v13i0.1811

Hübscher, I., Esteve-Gibert, N., Igualada, A., & Prieto, P. (2017). Intonation and gesture as bootstrapping devices in speaker uncertainty. *First Language*, *37*(1), 24–41. https://doi.org/10.1177/0142723716673953

Hübscher, I., & Prieto, P. (2019). Gestural and Prosodic Development Act as Sister Systems and Jointly Pave the Way for Children's Sociopragmatic Development. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/article/10.3389/fpsyg.2019.01259

Hübscher, I., Vincze, L., & Prieto, P. (2019). Children's Signaling of Their Uncertain Knowledge State: Prosody, Face, and Body Cues Come First. *Language Learning and Development*, *15*(4), 366–389. https://doi.org/10.1080/15475441.2019.1645669

Hudson Kam, C. L., & Edwards, N. A. (2008). The use of uh and um by 3- and 4-year-old native English-speaking children: Not quite right but not completely wrong. *First Language*, *28*(3), 313–327. https://doi.org/10.1177/0142723708091149

Jansson-Verkasalo, E., Silvén, M., Lehtiö, I., & Eggers, K. (2021). Speech disfluencies in typically developing Finnish-speaking children – preliminary results. *Clinical Linguistics & Phonetics*, *35*(8), 707–726. https://doi.org/10.1080/02699206.2020.1818287

Kim, S., Paulus, M., Sodian, B., & Proust, J. (2016). Young Children's Sensitivity to Their Own Ignorance in Informing Others. *PLOS ONE*, *11*(3), e0152595. https://doi.org/10.1371/journal.pone.0152595

Kools, J. A., & Berryman, J. D. (1971). Differences in Disfluency Behavior Between Male and Female Nonstuttering Children. *Journal of Speech and Hearing Research*, *14*(1), 125–130. https://doi.org/10.1044/jshr.1401.125

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*(4), 609–639. https://doi.org/10.1037/0033-295X.100.4.609

Krahmer, E., & Swerts, M. (2005). How Children and Adults Produce and Perceive Uncertainty in Audiovisual Speech. *Language and Speech*, *48*(1), 29–53. https://doi.org/10.1177/00238309050480010 20

Laserna, C. M., Seih, Y.-T., & Pennebaker, J. W. (2014). Um . . . Who Like Says You Know: Filler Word Use as a Function of Age, Gender, and Personality. *Journal of Language and Social Psychology*, *33*(3), 328–338. https://doi.org/10.1177/0261927X14526993

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods*, *41*(3), 841–849. https://doi.org/10.3758/BRM.41.3.841

Lipko, A. R., Dunlosky, J., & Merriman, W. E. (2009). Persistent overconfidence despite practice: The role of task experience in preschoolers' recall predictions. *Journal of Experimental Child Psychology*, *103*(2), 152–166. https://doi.org/10.1016/j.jecp.2008.10.002

Lyons, K. E., & Ghetti, S. (2010). Metacognitive Development in Early Childhood: New Questions about Old Assumptions. In A. Efklides & P. Misailidi (Eds.), *Trends and Prospects in Metacognition Research* (pp. 259–278). Springer US. https://doi.org/10.1007/978-1-4419-6546-2_12

Lyons, K. E., & Ghetti, S. (2013). I Don't Want to Pick! Introspection on Uncertainty Supports Early Strategic Behavior. *Child Development*, *84*(2), 726–736. https://doi.org/10.1111/cdev.12004

Maclay, H., & Osgood, C. E. (1959). Hesitation Phenomena in Spontaneous English Speech. *WORD*, *15*(1), 19–44. https://doi.org/10.1080/00437956.1959.11659682

Mele, C. (2017, February 24). So, um, how do you, like, stop using filler words? *The New York Times*. *https://www.nytimes.com/2017/02/24/us/verbal-ticks-like-um.html*

Paulus, M., Proust, J., & Sodian, B. (2013). Examining implicit metacognition in 3.5-year-old children: An eye-tracking and pupillometric study. *Frontiers in Psychology*, *4*. https://www.frontiersin.org/article/10.3389/fpsyg.2013.00145

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Podlesskaya, V. I. (2010). *Parameters for typological variation of placeholders*. Tsl.93.02pod; John Benjamins Publishing Company. https://benjamins.com/catalog/tsl.93.02pod

Preston, D. R. (1986). Five visions of America. *Language in Society*, *15*(2), 221–240. https://doi.org/10.1017/S0047404500000191

Richardson, E., & Keil, F. C. (2022). Thinking takes time: Children use agents' response times to infer the source, quality, and complexity of their knowledge. *Cognition*, *224*, 105073. https://doi.org/10.1016/j.cognition.2022.105073

Schneider, W. (1998). Performance prediction in young children: Effects of skill, metacognition and wishful thinking. *Developmental Science*, *1*(2), 291–297. https://doi.org/10.1111/1467-7687.00044

Schober, M. F., & Brennan, S. E. (2003). Processes of Interactive Spoken Discourse: The Role of the Partner. In *Handbook of Discourse Processes*. Routledge.

Shatz, M., Wellman, H. M., & Silber, S. (1983). The acquisition of mental verbs: A systematic investigation of the first reference to mental state. *Cognition*, *14*(3), 301–321. https://doi.org/10.1016/0010-0277(83)90008-2

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*(1), 25–38. https://doi.org/10.1006/jmla.1993.1002

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106

Streeck, J. (1996). A little Ilokano grammar as it appears in interaction. *Journal of Pragmatics*, *26*(2), 189–213. https://doi.org/10.1016/0378-2166(96)00012-4

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, *30*(4), 485–496. https://doi.org/10.1016/S0378-2166(98)00014-9

Tian, Y., Maruyama, T., & Ginzburg, J. (2017). Self Addressed Questions and Filled Pauses: A Cross-linguistic Investigation. *Journal of Psycholinguistic Research*, *46*(4), 905–922. https://doi.org/10.1007/s10936-016-9468-5

Tottie, G. (2011). Uh and Um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics*, *16*(2), 173–197. https://doi.org/10.1075/ijcl.16.2.02tot

Tottie, G. (2014). On the use of uh and um in American English. *Functions of Language*, *21*(1), 6–29. https://doi.org/10.1075/fol.21.1.02tot

van Amelsvoort, M., Joosten, B., Krahmer, E., & Postma, E. (2013). Using non-verbal cues to (automatically) assess children's performance difficulties with arithmetic problems. *Computers in Human Behavior*, *29*(3), 654–664. https://doi.org/10.1016/j.chb.2012.10.016

van Loon, M., de Bruin, A., Leppink, J., & Roebers, C. (2017). Why are children overconfident? Developmental differences in the implementation of accessibility cues when judging

concept learning. *Journal of Experimental Child Psychology*, *158*, 77–94. https://doi.org/10.1016/j.jecp.2017.01.008

Visser, M., Krahmer, E., & Swerts, M. (2014). Children's Expression of Uncertainty in Collaborative and Competitive Contexts. *Language and Speech*, *57*(1), 86–107. https://doi.org/10.1177/0023830913479117

Walker, E. J., Risko, E. F., & Kingstone, A. (2014). Fillers as Signals: Evidence From a Question–Answering Paradigm. *Discourse Processes*, *51*(3), 264–286. https://doi.org/10.1080/0163853X.2013.862478

Watanabe, M., Hirose, K., Den, Y., & Minematsu, N. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication*, *50*(2), 81–94. https://doi.org/10.1016/j.specom.2007.06.002

White, K. S., Nilsen, E. S., Deglint, T., & Silva, J. (2020). That's thee, uuh blicket! How does disfluency affect children's word learning? *First Language*, *40*(1), 3–20. https://doi.org/10.1177/0142723719873499

Yairi, E. (1981). Disfluencies of Normally Speaking Two-Year-Old Children. *Journal of Speech, Language, and Hearing Research*, *24*(4), 490–495. https://doi.org/10.1044/jshr.2404.490

Yairi, E., & Clifton, N. F. (1972). Disfluent Speech Behavior of Preschool Children, High School Seniors, and Geriatric Persons. *Journal of Speech and Hearing Research*, *15*(4), 714–719. https://doi.org/10.1044/jshr.1504.714

Zandan, N. (2018, August 1). How to Stop Saying "Um," "Ah," and "You Know." *Harvard Business Review*. https://hbr.org/2018/08/how-to-stop-saying-um-ah-and-you-know

Appendix

Table 10

Practice questions by Question Type and Difficulty Bin. "Answer" column indicates the accepted answer for each question. Accuracy rates are across children. Accuracy rates are binned by ratio for the numerical comparison questions.

Question Type	Question	Answer	Difficulty Bin	Accuracy Rate
Age	How old are you?	NA	Easy	100.00%
Age	How old is this person?	NA	Hard	0.00%
Numerical Comparison	Which side has more dots? (6 vs. 24)	24	Easy	96.21%
Numerical Comparison	Which side has more dots? (12 vs. 18)	18	Mid	89.39%
Numerical Comparison	Which side has more dots? (15 vs. 15)	Neither	Hard	0.00%
Animal ID	What kind of animal is this?	Groundhog	Hard	10.77%
Animal ID	What kind of animal is this?	Squirrel	Easy	91.04%
Animal Fact	What sound does a lion make?	Roar	Easy	91.04%
Animal Fact	What sound does a hyena make?	Hehehe	Mid	28.13%

Table 11

Test questions by Question Type and Difficulty Bin. "Answer" column indicates the accepted answer for each question. Accuracy rates are across children. Accuracy rates are binned by ratio for the numerical comparison questions.

Question Type	Question	Answer	Difficulty bin	Accuracy
Animal Fact	What are baby dogs called?	Puppies	Easy	75.76%
Animal Fact	What are baby cats called?	Kittens	Easy	77.27%
Animal Fact	What animal has the longest neck in the world?	Giraffe	Easy	79.10%
Animal Fact	What do cows eat?	Grass	Easy	81.18%
Animal Fact	What color is a pig?	Pink	Easy	84.85%
Animal Fact	What color is a dolphin?	Grey	Easy	55.22%
Animal Fact	What sound does a horse make?	Neigh	Easy	75.00%
Animal Fact	What sound does a cat make?	Meow	Easy	89.55%
Animal Fact	What are baby swans called?	Cygnets	Hard	0.00%
Animal Fact	What are baby llamas called?	Cria	Hard	1.61%
Animal Fact	What animal has the longest tongue in the world?	Anteater	Hard	13.43%
Animal Fact	What color is an octopus' blood?	Blue	Hard	26.87%
Animal Fact	What color is a hippo's milk?	Pink	Hard	16.42%
Animal Fact	What do blue whales eat?	Krill	Hard	16.42%
Animal Fact	What sound does a zebra make?	Heehaw	Hard	6.25%
Animal Fact	What sound does a cheetah make?	Chirping	Hard	7.81%
Animal Fact	What are baby deer called?	Fawns	Mid	7.81%
Animal Fact	What are baby sheep called?	Lambs	Mid	38.46%
Animal Fact	What animal has the biggest ears in the world?	Elephant	Mid	60.00%
Animal Fact	What color is a polar bear's skin underneath it's fur?	Black	Mid	43.28%
Animal Fact	What color is a robin's egg?	Blue	Mid	34.33%
Animal Fact	What do koalas eat?	Eucalyptus leaves	Mid	59.70%
Animal Fact	What sound do fish make?	Glub glub	Mid	68.66%
Animal Fact	What sound does a dolphin make?	Whistle/ clicking	Mid	48.48%
Animal ID	What kind of animal is this?	Snake	Easy	100.00%
Animal ID	What kind of animal is this?	Frog	Easy	98.53%
Animal ID	What kind of animal is this?	Rabbit	Easy	100.00%
Animal ID	What kind of animal is this?	Dog	Easy	97.01%
Animal ID	What kind of animal is this?	Pangolin	Hard	10.45%
Animal ID	What kind of animal is this?	Capybara	Hard	17.91%
Animal ID	What kind of animal is this?	Axolotl	Hard	14.93%

Animal ID	What kind of animal is this?	Elephant shrew	Hard	7.46%
Animal ID	What kind of animal is this?	Ferret	Mid	12.12%
Animal ID	What kind of animal is this?	Hamster	Mid	63.24%
Animal ID	What kind of animal is this?	Platypus	Mid	
Animal ID	What kind of animal is this?	Hedgehog	Mid	73.13%
Numerical Comparison	Which side has more dots? (6 vs. 24)	24	Easy	95.45%
Numerical Comparison	Which side has more dots? (15 vs. 15)	Neither	Hard	3.03%
Numerical Comparison	Which side has more dots? (12 vs. 18)	18	Mid	88.21%

Table 12

Individual differences in the number of fillers, hedges, and non-answers.

Participant Number	N Fillers	N Hedges	N Non-Answers
1	15	4	5
2	34	7	0
3	2	3	1
4	9	23	3
5	11	0	0
6	20	2	4
7	15	16	0
8	33	4	1
9	4	1	3
10	47	13	2
11	8	2	0
12	2	0	0
13	13	0	3
14	14	21	6
15	36	2	0
16	30	6	8
17	24	19	8
18	3	8	1
19	22	21	2
20	26	23	1
21	30	24	32
22	13	0	7
23	36	4	3
24	0	4	6
25	1	0	17
26	11	16	1
27	13	4	6
28	8	21	0
29	6	4	0
30	2	1	3
31	19	1	0
32	20	4	6
33	70	20	6
34	20	0	0
35	36	0	0
36	5	7	0
37	42	2	0
38	0	1	14
39	2	4	0

40	19	8	4
41	6	0	3
42	24	18	6
43	21	3	0
44	27	0	3
45	3	0	1
46	5	4	7
47	12	0	0
48	24	0	11
49	11	0	3
50	22	0	0
51	2	12	1
52	30	15	9
53	38	2	8
54	1	2	0
55	1	4	22
56	9	3	2
57	9	4	4
58	22	16	2
59	5	11	2
60	2	0	6