Metacognitive measures as predictors of accuracy in children

Madeleine Philippa Ingham

Dissertation submitted for the degree of

Psychology by Research

Master of Science

School of Psychology

College of Life and Environmental Sciences

The University of Birmingham

September 2021

ABSTRACT

Children often witness or are victims of crimes and so are required to provide memory evidence in court. Age is often used as a proxy for memory accuracy, meaning that legal decision makers treat testimonies from young children as unreliable, resulting in miscarriages of justice. Across two experiments, we investigated if implicit metacognition measures (e.g., vocal and body gestures, response time, and decision to hide an answer) and explicit measures (e.g., confidence) can be used to better predict the memory accuracy of children between the ages of 4-8. Children encoded complex episodic events and completed a 2-alternative-forced-choice task, then completed two self-report scales to measure their level of uncertainty. Predictive measures of accuracy included confidence, response time, box sorting, hedges, and fillers. Confidence was the most informative predictor of memory accuracy for all ages, suggesting that explicit measures are more indicative of memory accuracy than implicit measures in children of this age range. Moreover, confidence was more predictive of memory accuracy than children's age, suggesting that confidence can offer more information about children's likely memory accuracy than children's age. As such, these findings suggest that children in this age range have good metacognitive ability when encoding a complex memory, and that explicit measures of metacognition (i.e., confidence) and some implicit measures (i.e., response time, box sorting, hedges, and fillers) appear to be useful in predicting accuracy for children as young as 4.

ACKNOWLEDGMENTS

First and foremost, I’d like to thank Dr Melissa Colloff and Professor Heather Flowe for being the most wonderful supervisors, and for their encouragement, motivation, and support over the past two years - I'm incredibly grateful to have been able to work with them. I’d like to thank Dr Mahmoud Elsherif for giving me so much of his time - be it to help me with stats, coding, or to share words of encouragement. I’d also like to thank Professor James Rockey for sharing his time and statistical knowledge, the team at OBS for being so kind and supportive, and Sarah Rockowitz for being a constant source of hilarity and motivation. Finally, I’d like to thank the members of the Applied Memory Lab, the research assistants who volunteered to code the videos, and all the brilliant kids who took part in the experiments.

TABLE OF CONTENTS

OVERVIEW

CHAPTER 1: GENERAL INTRODUCTION

CHAPTER 2: EXPERIMENT 1

CHAPTER 3: EXPERIMENT 2

CHAPTER 4: GENERAL DISCUSSION

List of figures

List of tables

Appendices

OVERVIEW

Children have typically been deemed as unreliable witnesses in the Criminal Justice System (CJS). This is because age is often used as a metric to determine memory accuracy. However, a better metric may be to use metacognitive measures, such as confidence judgements (Mickes, 2015). This is because an individual with good metacognitive ability will be able to monitor when their memories are accurate or inaccurate. Whilst previously being regarded as a poor measure of accuracy, more recent research, and better methods of analysis (e.g., Winsor et al., in press) have shown that confidence is a better indicator of accuracy in adults and children than previously believed (e.g., Mickes, Hwe, Wais, & Wixted, 2011; Wixted et al, 2015; Wixted & Wells, 2017).

Eyewitness literature suggests that children aged 12 years and over can use confidence scales, and that children younger than 12 years do not yet have the metacognitive ability to monitor their memory or make accurate confidence judgements (Powell, Garry, & Brewer, 2013). However, developmental literature suggests that children aged 7 years also have a good confidence-accuracy relationship when the appropriate analysis (e.g., confidence-accuracy characteristic analysis) is used, and that children younger than 7-years-old have monitoring abilities but have trouble expressing this explicitly through confidence judgments (e.g., Lipko, Dunlosky, & Merriman, 2009; Lipko, Dunlosky, Lipowski, & Merriman, 2012; Roebers C. M., 2002; Winsor, et al., in press). The discrepancies in literature may exist for several reasons. First, tasks in eyewitness studies may be harder for children. Second, inappropriate analysis techniques may misrepresent children's metacognitive ability. Third, the eyewitness literature has not considered implicit measures as indicators of metacognitive ability. Some developmental research suggests that the issue with metacognitive ability may not be developmental, but methodological: children are developing metacognitive ability, but the methods of measurement used in eyewitness literature do not reflect this (e.g., Keast, Brewer & Wells, 2007; Powell, Garry, & Brewer, 2013). As such, children who may be too young to use map their uncertainty onto confidence scales may be able to monitor and indicate their memory implicitly, such as through vocal and body gestures, response time and the choice to hide an answer (e.g., Harris, Bartz, & Rowe, 2017; Hembacher & Ghetti, 2014; Kim, Paulus, Sodian, & Proust, 2016; Peterson & Briggs, 2001).

Whilst confidence has been explored in a forensic context, only of late has it been deemed a reliable indicator of accuracy in children (see Winsor, et al., in press, for a review). In contrast, there is little research on the utilisation of implicit metacognitive measures as predictors of accuracy in eyewitnesses. This thesis will investigate which implicit and explicit metacognitive measures best predict accuracy in neurotypical children aged 4-8 years, and consider how, with further research, these measures could potentially be utilised for application in a Criminal Justice Setting.

CHAPTER 1

GENERAL INTRODUCTION

1.1 Children in the Criminal Justice System

Children often witness or are victims of crimes and are thus required to provide memory evidence in criminal justice settings. 24,894 cruelties to children offences[1] were recorded in England and Wales in 2020/2021 (Statista Research Department, 2021), and as of March 2020, 51,510 children in the UK were in need of a child protection plan (Office for National Statistics, 2020). The number of child witnesses required to give memory evidence in the Criminal Justice System (CJS|) increased by 60% over a three-year period (2006-09; Plonikoff & Woolfson, 2011). The UK currently has no lower age limit for children being interviewed as a witness (Crown Prosecution Service, 2019), with children as young as 2-years-old having been called to give memory evidence in a legal setting (Bowcott, 2017; Plonikoff & Woolfson, 2011; *R v. Barker,* 2010). Despite this, the proportion of child abuse cases being closed due to 'evidential difficulties' remains high, at 49% in 2019 (Office for National Statistics, 2019), suggesting that issues and misconceptions in gathering evidence from young children still exist in the legal system. The Rochdale child sex abuse ring is a notable case of evidential difficulty hindering prosecution, where consistent and mutually supportive evidence of sexual abuse was continuously ignored because the witnesses—aged 13 years and older—were regarded as unreliable. As a result, the perpetrators were not convicted until almost a decade after the abuse had started (Carter, 2012; Smith, 2013). This is a concerning sentiment, as if children aged 13 years are believed to be unreliable, then it is

[1] Cruelty to children includes the following separate offences:
- Cruelty to and neglect of children
- Abandoning a child under the age of two years
(Statista Research Department, 2021)

possible that memory evidence from children younger than 13 will be given limited weight in court, and possibly resulting in miscarriages of justice (Pidd & Dodd, 2020).

1.1.1 History of the child witnesses

The belief that children are poor eyewitnesses assumes that age is an appropriate proxy for memory accuracy. Children are believed to have poor memory performance and awareness of when their memories are inaccurate. As a result, children have historically been regarded as unreliable and uncreditable witnesses by lay people (Goodman & Michelli, 1981; Katz & Mazur, 1979; Yarmey & Jones, 1983; Ceci, Ross, & Toglia, 1989; Wigmore, 1935/1976) and by legal professionals (e.g., Brainerd & Reyna, 2012; Knutsson & Allwood, 2014; Melinder, Goodman, Eilertsen, & Magnussen, 2004; Brigham & WolfsKeil, 1983; Featherstone & Kaladelfos, 2016) with early researchers (e.g., Binet, 1990; Stern, 1910; Whipple, 1909) appearing to have found ample evidence for children having poor recollection of actual events (McGough, 1993).

During the mid-20th century, prior to trial, jurors were warned of children's unreliability as witnesses, and to regard their evidence with caution (Hamilton & Addison, 1947). Lord Goddard extended this notion in 1958, condemning the calling of children under the age of 6 years with the belief that the court 'could not attach any value to the evidence of a child of 5' (*R. v Wallwork*, 1958). A 'truth vs lies' test was developed, where children were 'tested' prior to trial and considered to be even less reliable if they were not able to convey understanding of truth telling and the consequences of lying to a judge – a test that remains in practice at present (Rozell, 1985). Some states in the USA have implemented evidentiary corroboration rules for when a child is testifying. This requires the child's testimony to be confirmed by at least one other person (Ceci, Ross, & Toglia, 1987), and means that many testimonies were disregarded on the basis that young children were unable to give sworn evidence (Wade, 1997). Children were (and still are) often subjected to leading questions that infer an answer and challenges to their statements, both of which have been shown to contaminate memory evidence in children and adults (see Ceci, Hritz, & Royer, 2016 for review). The memory errors children make because of these poor practices may serve to further perpetuate the idea that children can be unreliable witnesses.

In more recent years, in England and Wales, special measures and guidelines have been introduced for when child witnesses are called to give memory evidence to improve the quality of the evidence and the experiences of the witness (Crown Prosecution Service, 2021). These measures have been met with both praise and criticism. The Youth Justice and Criminal Evidence Act (YJCEA) 1988 featured the abolition of the corroboration rule, with subsequent access to child eyewitnesses resulting in a substantial increase in the number of child sexual abuse prosecutions in the 1980s (Bell, 1986; Ceci & Bruck, 1995; Wade, 1997). The use of live video broadcasting was also introduced in place of the victim being present in the courtroom in the hope that this would reduce stress on child witnesses. However, the National Working Group and Victim Support (2014) identified occasions where there were failures in implementing special measures, such as screens being forgotten and video broadcasting links not working. These incidents often result in the child having to testify in the court room and *risk the child presenting as unreliable to the jury*, as they are visibly shaken by having to suddenly testify in person (Home Affairs Committee, 2013).

Intermediaries were introduced in the late 1990s with the YJCEA to relay information to and from both the witness and those involved in the legal process. It was hoped that this would protect children during cross examination (Cooper & Mattison, 2017), and that building rapport with the child would facilitate communication and memory disclosure

(Almerigogna, Ost, Bull, & Akehurst, 2007; Collins, Harker, & Antonopoulos, 2017; Plotnikoff & Woolfson, 2007; Ridley et al, 2015). Indeed, one study has shown that children with an intermediary provided more correct details during a recall task than children without an intermediary (Henry et al., 2017). However, many argued that communicating via an intermediatory could hinder the process of collecting memory evidence, as they believed information could be lost or misrepresented during the passing of information, and children may feel conflicted when being challenged by someone they had built rapport with (Davies, Hanna, Henderson, & Hand, 2011).

The YJCEA 1999 quotes under section 53 that children's baseline competence should be believed to be at the same level as that of adults. The Achieving Best Evidence guide was introduced to aid in interviewing vulnerable witnesses and advises that the development of a child must be considered when probing for certain details (e.g., time and date estimates), as it may be that only older children understand these concepts (Ministry of Justice, 2011). Although there are no longer restrictions on the age of a child witness testifying in court, beliefs towards and the practice of involving children in the CJS remain unstandardised. A study by Melinder and colleagues (2004) examined how legal professionals regard child witnesses and signified a large divide amongst legal professionals and their beliefs about child witness reliability, with defence attorneys being the most sceptical, and police the least. These results suggest that, despite advances in how children are treated in the CJS, discrepancies in practice and beliefs about children as eyewitnesses persist (e.g., Melinder, Goodman, Eilertsen, & Magnussen, 2004).

Given the number of children giving memory evidence in the CJS, it is integral that those involved in legal decision making are aware of the current knowledge about children's memory, and that children's memory evidence is collected and interpreted appropriately and consistently to ensure the outcomes of cases are fair (Bull, 2011). One possible way to determine the how much trust to place in witness memory evidence is to use metacognitive measures, such as confidence judgements (Mickes, 2015).This is because if a person has good metacognitive ability, they will know when they are underperforming on a task and will be able to indicate this uncertainty through confidence judgements (Brewer & Wells, 2006; Sauer, Brewer, Zweck, & Weber, 2010; Fleming & Lau, 2014; Winsor, et al., in press). Whilst this has been considered at length for adult witnesses (see Wixted & Wells, 2017 for a review), it has not been considered in close detail for child witnesses.

1.2 What is metacognition?

Metacognition refers to an individual's knowledge about and ability to control and monitor cognitive activity (Flavell & Wellman, 1997). John Flavell—a researcher at the forefront of metacognition—broadly defines the concept as 'knowledge of one's own cognitive processes, or anything related to them' (pp. 231-235, Flavell, 1976). Metacognition differs from cognition in that it has the added level of control over knowledge after knowledge acquisition (Vygotsky L. S., 1962). Monitoring and control are key aspects of the metacognitive framework: metacognitive monitoring involves feelings of uncertainty, whilst metacognitive control involves the implementation of strategy (Koriat & Goldsmith, 1996; Nelson & Narens, 1990). Combined, the two cooperate to direct and guide individual's problem solving (Fisher, 1998).

It is typically believed that metacognitive abilities are evident in children aged 6-8 years (e.g., Beck, Robinson, & Freeth, 2008; Metcalfe & Finn, 2013; Robinson & Whitaker, 1985; Roebers, von der Linden, & Howie, 2007) and robust by the age of 12 years (e.g., Kuhn, 1999; Powell, Garry, & Brewer, 2013; Pressley, Levin, & Ghatala, 1984), with many researchers asserting that metacognitive skills, such as monitoring and control, are only present in middle childhood (e.g., Beck & Robinson, 2001; Flavell, Green, & Flavell, 2000; Pillow & Anderson, 2006). Eyewitness literature has been particularly sceptical of children's metacognitive ability in the past, often concluding that children under the age of 12 years are unable to demonstrate monitoring and control abilities (Powell, Garry, & Brewer, 2013). This is possibly because tasks in eyewitness literature often measure metacognition in ways that mismeasure younger children's ability. For example, eyewitness researchers typically measure metacognitive monitoring explicitly through confidence judgements. Younger children have been shown to have difficulty mapping their uncertainty onto Likert-type scales typically used for confidence ratings (e.g., Chambers & Johnston, 2002; Powell, Garry, & Brewer, 2013). Similarly, metacognitive control is often measured using difficult tasks with complex linguistic demands that younger children lack (Darnell, 2015; Smith, Shields, & Washburn, 2003; Pratt & Bryant, 1990).

However, reanalysis of witness literature using more appropriate analysis techniques (e.g., confidence-accuracy characteristic analysis) demonstrates a strong relationship between confidence and accuracy (see Winsor, et al., in press, for a review). Newer eyewitness literature (e.g., Bruer, Fitzgerald, Price, & Sauer, 2017) has demonstrated that children aged 6-13 years were able to appropriately use confidence scales to provide information about their recognition memory (i.e., rate accurate answers as high confidence, and inaccurate answers as low confidence) Windsor et al., (in press) found in their study using a complex applied task that children from age 10 years had a good confidence-accuracy relationship, and children from age 7 years had an emerging relationship.

Additionally, there is evidence outside of eyewitness literature that suggests these metacognitive skills are present in early childhood (e.g., Balcomb & Gerken, 2008; Ghetti, Hembacher, & Coughlin, 2013). Developmental literature has indicated that when simple memory tasks are used, children as young as 4 seem able to appropriately rate their confidence. Hembacher & Ghetti (2014) found that 4 and 5-year-olds could appropriately rate their confidence for inaccurate and accurate answers in a memory retrieval task. They also found that low performing 3-year-olds had lower overall confidence, suggesting that whilst they cannot utilise the confidence scale at item level, they do have some awareness of their low performance. Lyons & Ghetti (2011) drew a similar conclusion in a perceptual discrimination task, with 3, 4- and 5-year-old's confidence judgments discriminating between accurate and inaccurate responses. The relevance of these findings and the reanalysis of previous data is that younger children may be unable to use confidence scales as well as older children for *complex tasks but* may still be aware that their memories are inaccurate. Using simpler, age-appropriate perceptual tasks, developmental literature (e.g., Balcomb & Gerken, 2008; Lyons & Ghetti, 2011; Lyons & Ghetti, 2013) recognises that metacognition may be present in children as young as 4, but that confidence judgements may not be the best way to represent this ability when tasks are complex. Together, this suggests that the issue with metacognitive ability is not always developmental, but sometimes methodological: younger children have metacognitive ability, but the methods of measurement used in eyewitness literature do not reflect this. Rather, younger children's metacognition may be better measured *implicitly* (e.g., Nilsen, Graham, & Chambers, 2008) rather than *explicitly*. The following section will underpin the main interpretation of metacognition and relevant terms to this thesis. It will also investigate developmental theories of metacognition and, specifically, monitoring and control.

1.2.1 An overview of metacognition

Flavell and his colleagues pioneered the research on metacognition (see Flavell, Miller, & Miller, 2002 for a review) and defined it as an individual's knowledge of their own cognition, knowledge about the nature of cognition, and knowledge about the skills and strategies related to cognitive activities. The theory was largely influenced by Piaget's work investigating children's ability to recognise that there can be various solutions to one problem (see Smith, 1994 for review). In his 1979 seminal paper, Flavell described a model of metacognition that comprised of four major components: metacognitive knowledge, metacognitive experiences, metacognitive goals, and metacognitive strategies.

Metacognitive knowledge involves knowledge and/or beliefs about an individual's cognitive processes; for example, an individual's knowledge about their own motivations, interests and abilities, and how external factors can affect this knowledge (Flavell & Wellman, 1977). It also includes knowledge about other's cognitive states, tasks, and strategies (Crescenzi, 2016). Metacognitive knowledge can be further divided into declarative and procedural knowledge. Flavell introduced declarative knowledge as "knowing what" (Flavell & Wellman, 1977; Schraw & Moshman, 1995). Following Flavell's definition of declarative knowledge, Brown (1987) introduced procedural knowledge which involves "knowing *how"*.

Metacognitive experiences are awareness and feelings elicited during a cognitive task (Schneider & Artelt, 2010). They act as 'quality control' (pp. 908, Flavell, 1979) for cognitive processes, and include the application of previous knowledge and information as resources to aid in a cognitive task. This previous experience provides feedback on the individual's performance and progress on a task, and in turn builds on their metacognitive knowledge.

Metacognitive experiences also play a role in strategy activation (Flavell, 1979). For example, when studying for a test, a child may realise that they have not studied a particular item for long enough to be able to perform well, and so choose to revise the item more. If they perform poorly on the test, they may identify that their strategy was not optimal. As a result, they will alter their revision strategy for the next test.

Metacognitive goals involve the desired outcomes of a task. Completion draws on both knowledge and experiences in terms of an individual's prior knowledge and experiences with a task. Individuals with good metacognition will be able to accurately assess how well they will perform and thus the expected outcome based on their prior knowledge. Outcomes can include producing something, solving a problem, or improving knowledge (Flavell, 1979).

Finally, metacognitive strategies include the regulation and control of cognitive activity to achieve the desired goal (Flavell, 1979). They are implemented during a task to monitor how well it is being performed and how to improve. Novel tasks are likely to be monitored more stringently than a previously performed task. Whilst isolated concepts, these four components draw heavily on each other during cognitive tasks. An individual with good metacognition will be able to utilise all components appropriately.

1.2.2 Meta-representational and non-meta-representational metacognition

Aspects of metacognition have different interpretations under declarative and procedural knowledge. Using a declarative meaning, metacognition requires meta-representation and conceptual understanding of mental states (Esken, 2012). Assessments of metacognition based on the declarative interpretation of monitoring and control often involve verbal self-report, as individuals are assumed to have awareness of their knowledge state.

However, this often proves problematic for younger children who may have difficulty verbalising their epistemic state (Waters, 2009)

When considered under a procedural meaning, metacognition does not necessarily require meta-representation or conceptual understanding of mental states (Proust, 2007). Rather, this view asserts that aspects of metacognition can be demonstrated through behavioural sensitivity, such as feelings experienced during a cognitive task (Koriat & Levy-Sadot, 1999; Proust, 2012). Beran and colleagues (2013) use the example of a child who becomes anxious when faced with making a decision. Being in a state of anxiety indicates to their cognitive system that action must be taken to make the decision. A child's behavioural sensitivity to their state of knowledge is sufficient to inform metacognition without needing meta-representation.

Proust's (2010) 'two functions of self-knowledge' encompasses both non-meta representational and meta-representational views, in that implicit metacognition does not require conceptual understanding, but explicit metacognition does. Younger children can only demonstrate non-reflective, implicit metacognition, with monitoring and control being based on 'experience and feelings' (Proust, 2007). As children's metacognitive abilities develop, they move from being just able to demonstrate implicit metacognition to being able to demonstrate explicit metacognition too (Darnell, 2015).

In contrast, Perner's (2012) 'Minimally metacognitive' theory is in line with meta-representation, where implicit and explicit metacognition exist on a continuum 'from ordinary object-level cognition to full-blown recursive cognition' (pp. 97, Perner, 2012). Both are meta-representative, but implicit skills are a precursor to explicit skills (i.e., children first demonstrate implicit knowledge, and move towards explicit knowledge as their metacognition develops). Both require recognition of being in a state of knowing and thinking about the content of the knowledge. As such, younger children begin with just implicit skills and develop explicit skills later in life (Geurten & Bastin, 2018).

1.2.3 Explicit and Implicit metacognition

Explicit metacognition is inherently meta-representative, as it involves working consciousness and an individual’s ability to actively reflect on and report their uncertainty (Smith, Shields, & Washburn, 2003). Explicit skills are considered a ‘reflective’ form of metacognition and occur at an age when children can actively reflect on their uncertainty. Explicit behaviours can include verbal expressions of epistemic stance, like confidence judgements (Esken, 2012). Children who can report their uncertainty explicitly are able to demonstrate verbal and behavioural awareness of their ignorance (Darnell, 2015). Children may be able to monitor and verbally report their uncertainty using confidence judgements and use this knowledge of their epistemic stance to influence their subsequent strategic behaviour. Kloo & Rohwer (2012) and Schneider & Lockl (2002) state that only from age 6 years are children able to appropriately verbalise their uncertainty explicitly, although Gopnik & Graf (1988) found that 5-year-olds could accurately identify the source of their knowledge when asked how they knew the answer to something, indicating an explicit and reflective awareness of knowledge states.

Implicit metacognition describes metacognitive processes being driven by behavioural sensitivity, or in Perner’s words, ‘*being* in a state’ rather than ‘*knowing* one is in this state’ (pp. 98, Perner, 2012). Although young children may not yet demonstrate overt verbal or behavioural awareness to uncertainty, they may be able to monitor and guide their behaviour through access to implicit knowledge (Balcomb & Gerken, 2008; Wellman, 1977). Implicit behaviours include vocalisations, like fillers and hedges (e.g., fillers: ‘um’, ‘hmm’, hedges: ‘could be’, ‘maybe’; Esposito, Marinaro, & Palombo, 2004; Gustafsson, Lindholm, & Jönsson, 2019); body gestures, like head shakes and shrugs (e.g., Harris, Bartz, & Rowe, 2017; Debras, 2017), response time (e.g., Patterson, Cosgrove, & O’Brien, 1980), and answer skipping or withholding (e.g., Balcomb & Gerken, 2008; Hembacher & Ghetti, 2014). Children may be sensitive to their own uncertainty, and act in accordance with a state of ignorance (Darnell, 2015).

Under a non-metarepresentational view, implicit metacognition ‘guides behaviour but does not reach conscious awareness’ (pp. 89, Brinck & Liljenfors, 2013) and involves pre-reflective knowledge (Kloo & Rohwer, 2012; Reder & Schunn, 1996). Implicit metacognitive skills are believed to represent a ‘pre-reflective’ form of metacognition. In their study, Gopnik & Graf (1988) found that 3-year-olds could indicate they knew something but could not identify their knowledge source. Darnell (2015) found that 5- and 6-year-olds behaviour was not influenced by the ease with which they can produce an answer, indicating a sensitivity to their knowledge state. Call & Carpenter’s (2001) study showed that children from the age of two would check for an item at different locations before commitng to a chocice when they had not seen the item being hidden, indicating a sensitivity to their uncertainty about their knowledge state.

Under a meta-representational view, implicit metacognition is a precursor to explicit metacognition and exists as a sensitivity to alternatives (Perner, 2012). Referring again to Call & Carpenter’s (2001) study as an example, Perner described that if children continue to check the different locations to see where the item had been placed before committing to a choice, then this is evidence of reflecting upon their ignorance. In other words, the children’s continued searching behaviour indicates an awareness of and sensitivity to possible alternatives (Darnell, 2015).

1.2.4 The development of metacognition

In developmental literature, early implicit metacognitive abilities appear to exist in children's second year of life. Error detection has been demonstrated in 18-month-olds by Poulin-Dubois and colleagues (2007), who found that children looked longer when an adult searched for an item at a new location despite having been blindfolded during the transferring of the item. Brown (1983) notes that metacognitive skills develop slowly during preschool years, with younger children demonstrating less refined metacognitive skills and older children being better at learning and performing certain tasks due to more refined metacognitive awareness (Flavell, Green, & Flavell, 1995)

There are various theories for the development of metacognition. Flavell believed that metacognitive ability changes with age, asserting that development is aided by experience, and that the trajectory of development depends on different learning contexts and opportunities (Flavell, Green, & Flavell, 1995). When children are provided with the tools and opportunity to utilise metacognitive skills, they are likely to develop sound metacognition.

Vygotsky believed that social interaction with others (particularly adults, as they already have refined metacognitive skills) facilitates the gradual internalisation of self-regularity skills, which include an individual's thoughts and actions (Vygotsky L. S., 1962). Discussion with others is helpful in exploring new ideas and interpretations (Harry-Augstein & Thomas, 1991). Metacognitive skills exist initially in an interpsychological (i.e., social) context, with interactions with others helping move these skills to an intrapsychological (i.e., individual) context (Papaleontiou-Louca, 2003). Language largely facilitates this by aiding in the mental abstraction of knowledge. This social facilitation happens naturally from an early age with the child's caregiver in the home environment and is furthered in schools. Vygotsky coined the term 'private speech' to describe the audible but non-directed commentary that young children engage in whilst carrying out tasks. Private speech has been identified in children as young as two, and seems to peak between 4-5, before slowly declining as self-commentary is internalised around ages 7-8 years (Vygotsky, 1962; Winsler, Fernyhough, & Montero).

Bruner (1978) used Vygotsky's work as a basis for his notion of 'scaffolding'. Like private speech, which involves the implementation of strategy and mediation by adults to aid metacognitive development in children. Communicating with others aids in developing children's cognitive processes and consciousness to a level where they can communicate them (Vygotsky, 1978). Conversation allows children to become more aware of their epistemic states (Kim, Paulus, Sodian, & Proust, 2016), or a 'set of beliefs', organised into theories and operating at the metacognitive level (Hofer, 2004).

Vygotsky also asserted that children's attention is initially controlled and considered in relation to others. Adults' direct children's attention to stimuli through verbal and physical hints (e.g., pointing). This learning from others may begin as early as age one, when children have been shown to follow an adult's gaze to a target object (Scaife & Bruner, 1975).

Brinck & Liljenfors (2013) further this notion, stating that the development of metacognition is rooted in intersubjectivity, or the sharing of experiences between child and caregiver. In contrast to Vygotsky's theory, Brinck & Liljenfors argue that interaction with an individual's environment and others continues as a part of their metacognition, and that metacognition is never fully independent of others and environment (Brinck & Liljenfors, 2013). Adults initiate this development in children by capturing and directing their attention in a dyadic–or a paired–interaction. Parental scaffolding aids in an infant's acquisition of knowledge. For example, infants have been shown to attend more to an object that is the subject of joint attention with an adult than an object that is not (Reid, Hoehl, & Striano, 2006). This suggests that infants are aware that adults can provide knowledge about an object (Campos & Stenberg, 1981; David & Appell, 1961; Feinman, 1992).

Referring to it as reflective abstraction, Jean Piaget believed that metacognition was related to knowledge and monitoring of others and the environment and developed through interaction with both (Piaget, 1964/1968). Specifically, he believed that individuals neither had nor needed access to their own internal knowledge; rather, their knowledge and self-regulation would be reflected in others and in the environment. Piaget believed that pivotal metacognitive development happened between the ages of 4-9 years, with children's awareness of themselves as learners developing through exposure to different viewpoints and when their understanding is challenged (Fisher, 1998).

1.2.5 Monitoring and control

Metacognitive monitoring involves an awareness and monitoring of cognition to inform and guide strategic behaviour for optimal performance of cognitive tasks (Brown, Bransford, Ferrara, & Campione, 1983; Gill, Swann, & Silvera, 1998; Nelson & Narens, 1990; Son & Schwartz, 2002). Monitoring involves several key functions: (1) ease of learning judgements (EOLs: predictions about how easy something will be to learn); (2) judgements of learning (JOLs: predictions about future performance); and (3) feelings of knowing (FOKs: judgements about whether information is available in memory, which in turn encompasses feelings of certainty and feelings of confidence (Koriat, 2000; Nelson & Narens, 1990). Metacognitive control involves the control of cognitive activities through a chosen strategy (e.g., Nelson & Narens, 1990). For example, when studying for an exam, an individual may monitor the difficulty of an item they are studying and control the allocation of study time in response to the level of difficulty. If an item is difficult, an individual with good metacognition with recognise through monitoring that their level of understanding is not optimal, and, through control, increase the time spent studying this item. Control also involves knowledge of recall readiness and strategy application, control of cognitive activity, and control of behaviour through strategy (Grainger, Williams & Lind, 2015). For example, the allocation of study time and when to terminate learning.

Nelson & Narens (1990) furthered Flavell's work and developed a framework that described three main principles of metacognition. Their work explained that logically, initiating metacognitive control initially requires monitoring, as information is gained during the monitoring of a task. In other words, an individual needs to have been monitoring initially to employ metacognitive control. Various other studies concur (e.g., (Nelson & Dunlosky, 1994). Monitoring and control can operate implicitly or explicitly. In other words, individuals can inform strategy through sensitivity to being in a state of ignorance (implicitly), or through actively reflecting upon their uncertainty (explicitly) (Kloo & Rohwer, 2012).

1.2.6 The development of monitoring and control

The developmental sensitivity of monitoring and control has been demonstrated in numerous studies (e.g., Hembacher & Ghetti, 2014; Lyons, 2011; (Metcalfe & Shimamura, 1994/1996; Scheider, Visé, Lockl, & Nelson, 2000). Istomina (1975) tasked children aged 4-7 years to buy items for a tea party using a shopping list that they were unable to take with them into the shop. Children aged 4 years would run back and forth between the shop and the list, 5 and 6-year-olds would ask the list to be repeated in attempt to memorise the items, and 7-

year-olds would try to logically connect the items. The difference in behaviour evidences the improvement in metacognitive skills with age: whilst the youngest children were sensitive to their uncertainty and able to monitor their level of knowledge as a result, as demonstrated by them returning to the list when they realised, they could not remember, older children seemed better able to use their memory monitoring to employ a strategy to help them remember the items. 5 and 6-year-olds appeared to employ a simple strategy of item repetition to aid in remembering, whilst the 7-year-olds strategy' using logical links suggests more sophisticated metacognitive ability.

1.3.8 Developmental differences of metacognitive ability

Why might younger children be poorer at expressing uncertainty explicitly using and implementing strategic behaviours compared to older children? The *availability deficiency* hypothesis posits that younger children lack metacognitive skills, and so cannot utilise them during relevant tasks (Veenman, Kerseboom , & Imthorn, 2000; Winne, 1996). Simply put, they do not know *how* to monitor their uncertainty and control their behaviour. However, it could be that they have metacognitive ability, but are not yet able to actively reflect upon it. The *production deficiency* hypothesis explains that younger children may have the ability to execute certain metacognitive skills but fail to spontaneously implement strategy when appropriate (Flavell, Beach, & Chinsky, 1966). For example, they may not understand the relevance of performing a particular strategy for the given task; in other words, they do not know *when* to monitor their uncertainty and control their behaviour. Proust's (2010) 'two functions of self-knowledge' posits that metacognition is a made up of 2 functions: (a) system 1: implicit and pre-reflective unconscious heuristics; and (b) system 2: an explicit and reflective conscious process. Under Prouts's idea, young children may be less adept at using explicit metacognition (system 2) because it involves meta-representation of cognitive states, which they are not able to do. Younger children may not be able to actively reflect upon their uncertainty, and so cannot verbally express it. Rather, they may rely mostly on implicit metacognition (systems 1). Donaldson (1978) furthers this notion by stating that to 'control and direct thinking' an individual 'must become conscious of it' (p.94, Donaldson 1978). To be able to control and direct the behaviour, children must become aware of their consciousness, indicating a transition from non-conceptual, implicit metacognition to conceptual explicit metacognition.

Under Perner's (2012) 'Mini-meta' idea, implicit metacognition may also be considered metarepresentational. It could be that young children have conceptual awareness of their mental state but are not able to use this information to implement explicit monitoring and control because their metacognitive skills are not yet as refined as older children's (Schneider & Lockl, 2002).

1.3 Summary

To summarise, it appears that both older and younger children can monitor their uncertainty but appear to experience and express this differently. This divide becomes more apparent when complex tasks are used in experiments (i.e., simple memory tasks in developmental literature, vs complex memory tasks in eyewitness literature). Older children may be able to *verbally express their uncertainty explicitly*, whilst younger children appear to be better able *to indicate uncertainty implicitly*. Considering control, older children seem more capable of implementing an appropriate strategy as informed by monitoring than younger children.

1.4 The confidence-accuracy relationship

1.4.1 Using confidence as predictor of memory accuracy

Confidence is often used as a proxy for accuracy in forensic settings. It has been an influential variable in evaluating witness accuracy, with the US Supreme Court being one of many legal systems to endorse its usage (e.g., Neil v. Biggers, 1972). Historically, legal professionals have valued a witness's testimony more highly if they indicate high confidence (e.g., Brigham & WolfsKeil, 1983; Brigham, 1990; Penrod & Cutler, 1990).

Lay people have also been shown to be influenced by confidence. Research has shown that eyewitness confidence is a strong predictor of the verdicts of mock jurors (Leippe, Manion, & Romanczyk, 1992). Jurors are more likely to assume a witness's answers are correct if they express high confidence in court (Brewer & Burke, 2002). Jurors have also been found more likely to advocate for a witness who reports high confidence compared to one who reports low confidence (Brewer & Burke, 2002). Cases where the witness is not 100 percent sure of their identification of a criminal suspect in a police lineup often results in the acquittal of the suspect (Wells, et al., 1998). A benefit of using confidence rather than age to determine a child's accuracy is that confidence can be collected for individual pieces of information, but age can only be used to provide a general proxy of likely accuracy. As witness statements contain a mixture of correct and incorrect information (Brown, et al., 2013; Memon, Meissner, & Fraser, 2010), information about the reliability of individual items would be useful for identifying which information to investigate further (e.g., statements made with high confidence).

1.4.2 The confidence-accuracy relationship in adults

Eyewitness researchers historically concluded a weak relationship between confidence and accuracy in adults. In a review of 31 studies, Wells & Murray (1984) made the concerning conclusion that "the eyewitness accuracy-confidence relationship is weak under good laboratory conditions and functionally useless in forensically representative settings", with an average correlation of 0.07 (Wells & Murray 1984, p. 165). Later reviews suggests that confidence-accuracy correlations average is 0.3 in adults; a slight increase but still a minimal correlation (e.g., Bothwell, Deffenbacher, & Brigham, 1987; Cutler, Penrod, & Martens, 1987b; Wells & Murray, 1983).

These findings were at odds with basic memory research, which has consistently found a strong confidence-accuracy relationship with adults (e.g., Juslin, Olsson, & Winman, 1996; Mickes, Hwe, Wais, & Wixted, 2011). Mickes, Wixted, & Wais, (2007) asked participants to rate their confidence on whether a present item was new in a recognition memory task. Their results demonstrated a strong confidence-accuracy relationship, with correct answers being rated as high confidence, and incorrect answers being rated as low confidence. More recently, Kurdi and colleagues (2017) asked participants to memorise and recall a list of words after a distractor task. The results showed a strong confidence-accuracy relationship for all items.

1.4.3 The confidence-accuracy relationship in children

Previously, eyewitness memory literature concluded that *children cannot provide explicit confidence judgements that reflect their accuracy* (Keast, Brewer & Wells, 2007; Powell, Garry, & Brewer, 2013). This is because they are believed to have poor metacognitive skills, and so cannot appropriately relate their certainty using confidence judgements. Keast

(2007) asked children aged 10-14 years to watch a mock crime video then identify the perpetrator. Their results suggested that the calibration between confidence and accuracy was poor, and so they concluded that children's confidence ratings provide no information on the guilt or innocence of a suspect. Brewer and Day (2005) asked children to identify a target suspect from a lineup after encoding a video event. They found that children were significantly overconfident in their judgements and concluded that confidence did not predict accuracy. The inference from these and other findings is that children may not yet have the metacognitive ability to monitor and indicate their certainty using confidence scales, often reporting overconfidence.

However, reanalysis of previous literature has suggested that children's confidence-accuracy relationship may have been misrepresented. For example, Keast, Brewer, & Wells (2007) concluded from their results that children had no confidence-accuracy relationship. When the results were reanalysed using CAC (see section 1.4.4.2) analysis, children from the age of 8 years had a good confidence-accuracy relationship, with younger children showing an emerging ability (see Winsor et al., for review). Additionally, more recent eyewitness research shows a positive picture of the confidence-accuracy relationship. For example, in one eyewitness identification study, Bruer, Fitzgerald, Price, & Sauer, (2017) asked children ages 6-13 years to watch a six-minute video alternating between a man (the target) reading a list of words and a woman performing a magic trick, and then later identify the man from a lineup. The authors collected children's confidence judgements using a water-cup rating scale, on which an empty cup indicated low confidence and a full cup indicated high confidence. Their results showed that confidence judgements reflected their accuracy at a group level. Similarly, Winsor et al. (in press) used the same water-cup scale in another eyewitness identification study sampling children in young- (4–6 years), middle- (7–9 years), and late- (10–17 years) childhood, and found that, after watching a video complex episodic event, children from the age of 7 years were able to use confidence to indicate how sure they were of their identification.

1.4.4 Confidence measuring techniques

There are various ways to measure confidence. The confidence-accuracy relationship has been shown to be unaffected by the number of points on a confidence scale (Dodson & Dobolyi, 2015a; Tekin & Roediger, 2017; Tekin, Lin, & Roediger III, 2018). Allwood, Granhag, & Jonsson, (2006) found children aged 11-12 performed equally on different scales (numerical, picture, line, and written scale), although the scales were supplemented with percentages. However, in terms of labelling, numerical polarity has been shown to effect individual's ratings, with higher ratings given on a scale range of -5-5 compared to a scale range of 0-10 (Schwarz et al., 1991). It could be that individuals do not want to ascribe their certainty to a negative value which may be viewed as an 'explicit failure' (Händel & Fritzsche, 2014). Händel & Fritzsche (2014) note that whilst verbal scales (e.g., 'not confident' – 'very confident') already have explicit meaning, non-verbal scales require personal interpretation, and will likely yield a higher level of variation between individuals. Some scales may suggest multiple meaning: use of symbols, such as smiley faces (Pressley et al., 1987; Roebers C. M., 2002) may imply an emotional component, such as how satisfied the individual is with their answer.

1.4.5 Confidence analysis techniques

The discrepancies in literature on confidence have been partially attributed to the fundamental differences in the analysis of the confidence-accuracy relationship. Juslin et al

(1996) asserted that the conclusion of a weak confidence-accuracy relationship in eyewitness literature stems from the use of an inappropriate form of analysis; namely, the point-biserial correlation coefficient, which is now deemed a controversial way to calculate the relationship between confidence and accuracy. The introduction of more appropriate analyses has revealed a stronger confidence-accuracy relationship. These methods of analysis are discussed in detail, next.

1.4.5.1 Point-biserial Correlation Coefficient

Early eyewitness literature often relied on the point-biserial correlation coefficient. The analysis involves correlating the responses for chooser (an individual who makes an identification from a lineup) and non-choosers (an individual who makes no identification from a line up: ‘not present’) with the corresponding confidence level producing a between-subjects correlation. This method of analysis can be misleading when considering the confidence-accuracy relationship: Juslin (1996) showed that the point-biserial correlation coefficient does not accurately represent the confidence-accuracy relationship, because even when the relationship between confidence and accuracy is perfectly calibrated (i.e., a witness chooses 50% confidence when they are 50% accurate), the correlation results can be skewed if the confidence ratings are not uniformly distributed. For example, the correlation results may be lower than expected if the distribution of confidence ratings are unimodal (i.e., clustered around the middle ratings), or higher than expected if the distributions of confidence ratings are bimodal (i.e., clustered at both extreme ratings).

1.4.5.2 Confidence-accuracy characteristic analysis

Confidence-accuracy characteristic (CAC) analysis involves plotting subjective confidence against objective performance, or proportion correct (Winsor, et al., in press).

Confidence scales for CACs can vary (e.g., a confidence scale of 100 or of 5; Mickes, Moreland, Clark, & Wixted, 2014). Research that has utilised this method has suggested a strong confidence-accuracy relationship in adults (e.g., Tekin, Wenbo, & Roediger III, 2018), and in children (e.g., Hiller & Weber, 2013). Considering eyewitness research, reanalysis of Keast, Brewer, & Wells (2007) lineup experiment demonstrated a strong confidence-accuracy relationship in both children (M age = 11) and adults (Winsor, et al., in press). Previous research that originally used a correlation coefficient to quantify the confidence-accuracy relationship has since also been reanalysed and shown that confidence correlates strongly with accuracy (Wixted, 2018). Reanalysis of developmental literature has also shown a strong confidence-accuracy relationship (see section 1.4.5). A continuous recognition task by Berch & Evans (1973) found that children aged 5-9 years had a good confidence-accuracy relationship, with older children's performance being slightly better. Finally, reanalysis of a facial recognition task by Wilkinson et al (2010) found a similarly strong confidence-accuracy relationship in neurotypically developing nine- to 17-year-olds. These recent conclusions are consistent with basic memory research, which has consistently found a strong confidence-accuracy relationship with adults (eg. Juslin, Olsson, & Winman, 1996; Kurdi, Diaz, Wilmuth, Friedman, & Banaji, 2017; Mickes, Hwe, Wais, & Wixted, 2011).

1.4.6 A confidence-accuracy characteristic reanalysis (Hembacher & Ghetti, 2014)

Developmental literature has indicated that children can accurately assign their confidence judgements from as young as 4 when simple memory tasks are used (Lyon & Ghetti, 2013). Hembacher and Ghetti (2014) investigated children's uncertainty monitoring during a simple memory task. They used different analysis than used in eyewitness literature: their analysis involved plotting the mean confidence for correctly and incorrectly identified items (a common analysis in the developmental literature). Whilst this was useful in demonstrating the difference in ages, it does not capture all the information as with a CAC analysis. To demonstrate this, we will consider the Hembacher and Ghetti (2014) study in more detail, next. Whilst this study is useful in evidencing young children's ability to monitor their uncertainty, these results do not indicate exactly how strong or weak the confidence-accuracy relationship is. By plotting the proportion correct for each level of confidence for this experiment (as in a CAC analysis), we can better visualise the confidence-accuracy relationship in children of this age group and during this task (Winsor et al., in press).

Hembacher and Ghetti used a 2-alternative-forced-choice (2AFC) object recognition task to demonstrate uncertainty monitoring in children aged 3-5 years. Children were shown 30 different line drawings, then asked to identify the already seen drawings from novel drawings. After choosing an answer, children were asked to rate their certainty on a three-point confidence scale with cartoon drawings of a child expressing various levels of certainty. A box sorting task was then introduced: children were asked to sort their answers into either an open-eye box or a closed-eye box and told that only the answers in the open-eye box would be checked to determine their final prize. The results indicated that children aged 5 years can accurately monitor and report their uncertainty and, on average, can give higher confidence ratings to their accurate answers and lower confidence ratings to their inaccurate answers. 4-year-olds were also capable of uncertainty monitoring, but the skill was more robust in the older children. Their results also suggested that whilst uncertainty monitoring was absent in 3-year-olds, there was evidence of emerging performance awareness, with high performing children being more confident.

Rather than looking at average confidence for accurate and inaccurate answers, we were interested in how accurate children were at each level of confidence. To better visualise the confidence-accuracy relationship, we reanalysed the data (obtained from the Open Science Framework) by plotting proportion correct for each level of confidence using CAC analysis. Proportion correct was calculated for each level of confidence for each age group (see Figure 1). The CAC plot shows that confidence-accuracy relationship matches the developmental trajectory of uncertainty monitoring, with the relationship changing from absent at 3 to stronger at 4 and best at 5. 5-year-olds were 93% accurate at the highest level of confidence, and 69% accurate at their lowest level. 4-year-olds were 86% and 68% accurate respectively, and 3-year-olds had similar accuracy for high (86%) and low (82%) ratings, suggesting they were unable to appropriately assign their confidence. As such, confidence is a strong indicator of overall memory accuracy in children aged 4-5 years.

This reanalysis indicates that children from age 4 years seem able to use confidence scales, and that 3-year-olds show an emerging ability. This could be evidence that, at least for simple memory tasks, confidence judgment from children as young as 4 seems to reflect their likely accuracy. This reanalysis could provide evidence for children as young as of 4 having good metacognitive ability and being able to utilise a confidence scale on a simple memory task. In sum, plotting proportion correct at each level of confidence provides us with more information about the correspondence between confidence and accuracy than if just average confidence is considered (as in Hembacher & Ghetti, 2014). Thus, we conclude that CAC analysis is the optimal technique in terms of visualising and understanding the confidence-accuracy relationship.

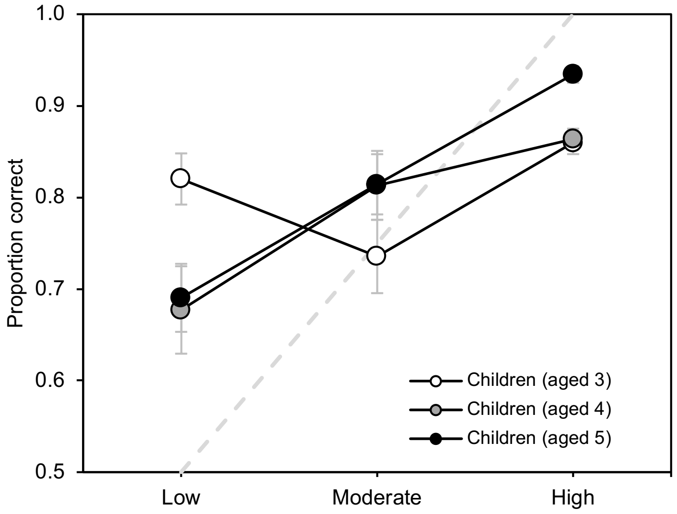


Figure 1. A CAC reanalysis of Hembacher & Ghetti (2014). The dashed line represents chance-level performance (Winsor, et al., in press).

1.4.7 Metacognitive measures as predictors of accuracy in children

Metacognition has been widely explored in developmental literature but not in the eyewitness literature. Rather, eyewitness literature has largely focused on *collecting explicit metacognitive measures using confidence judgments*. This is an unrepresentative conclusion of metacognitive ability, as often younger children cannot use confidence scales on eyewitness tasks. As such, eyewitness literature concludes that only children from the age of 12 years have good metacognitive ability, because only they seem able to use confidence scales (Powell, Garry, & Brewer, Eyewitness testimony, 2013). However, the developmental literature has shown that children under the age of 7 years have good metacognitive ability but are less able to indicate their uncertainty explicitly (Schneider & Lockl, 2002). Rather, they seem able to indicate their uncertainty implicitly (e.g., Balcomb & Gerken, 2008).

There appear to be three main reasons for these discrepancies in children's metacognitive ability in the eyewitness and developmental literature: 1) inappropriate analysis; 2) difference in tasks between literature; and 3) implicit metacognition being measured in developmental literature, but not in eyewitness literature. As with adults, data from research on confidence in children may have used inappropriate analysis, and so underestimated the confidence-accuracy relationship. Eyewitness literature has often used the Point-biserial Correlation Coefficient which, as discussed above, has been found to underestimate the confidence-accuracy relationship. Additionally, the difference between tasks in developmental and eyewitness literature may also yield different results. Developmental paradigms may be easier for children to execute than eyewitness tasks. This may account for why developmental literature generally finds a confidence-accuracy relationship in children younger than those in the eyewitness literature. Finally, eyewitness studies often require children to explicitly self-report their epistemic stance during or after a complex task using confidence judgements, which, as previously discussed, younger children may not be able to do. Relying on younger children's ability to self-report may also skew the representation of their metacognition ability (Winne & Perry, 2000), as children have more difficulty understanding the scales being used during the task (e.g., confidence scales), and may not have the verbal ability to use these scales. Reporting retrospectively can also be problematic: Flavell found that when children performed spontaneous verbal utterances, 25% of them were unable to recall doing so when asked to report what they had said (Flavell, Speer, Green, August, & Whitehurst, 1981). This suggests that a more optimal way to observe implicit behaviours in younger children may be through systematic observational studies.

From the evidence presented it appears that children younger than 7 can monitor the likely accuracy of their memories (Goupil, Romand-Monnier, & Kouider, 2016; Monosov & Hikosaka, 2013) but may have difficulty reliably indicating their epistemic stance explicitly through use of confidence judgements when complex tasks are used. As developmental research posits that implicit metacognition is evident in children around the age of 2 years, and explicit metacognition is evident in children aged 4-8 years (Sodian, Thoermer, Kristen, & Perst, 2012), we explore the developmental trajectory of metacognitive monitoring and control in children aged 4-8 years to better understand how these components develop are utilised. If children from age 4 years have good metacognitive awareness and uncertainty monitoring but are not yet reporting this explicitly on complex tasks, then utilising *implicit metacognitive measures* of uncertainty rather than *explicit metacognitive measures of uncertainty* may provide more information on their certainty, and therefore possibly memory accuracy. Considering these points, it could be that children are indeed better eyewitnesses than previously believed, as they can monitor their uncertainty from the age of 4 years and can give confidence judgements that accurately reflect their accuracy from at least age 7 years in eyewitness tasks (Winsor et al., in press).

How can we quantify certainty and uncertainty so they can be appropriately applied in a forensic context (i.e., a child giving evidence in court)? As described above, implicit metacognitive measures are behaviours that mark the epistemic stance of an individual. They are performed without being explicitly asked to report them. For example, when an individual is asked a question that they are uncertain of, they may shrug and state ‘I’m not sure’. They are explicitly stating their uncertainty through the ‘not sure’ statement, but also implicitly marking this uncertainty with a shrug.

1.5 Thesis Aim

The aim of this thesis is to connect the eyewitness and developmental literature, and to apply the developmental theory of metacognition to resolve a forensic problem. The discrepancies in literature on confidence have been attributed to the fundamental differences between eyewitness identification tasks and simple memory tasks, and how they are subsequently analysed. Nevertheless, there is promise that metacognition could be useful for determining reliability of memory evidence from children. What is currently unknown is whether implicit measures of metacognition can be used to predict memory accuracy on a complex episodic memory task in children aged between 4-8 years. Some recent research suggests that this may be the case. If children from the age of 7 years seem to be able to explicitly express their uncertainty through confidence scales, and children under 7 seem able to monitor an report their uncertainty implicitly, it follows that children between the ages of 4-7 years may be sensitive to their uncertainty, and able to express this implicitly (e.g., through body and vocal gestures), and children from the age of 7 years may be able to express their uncertainty explicitly (e.g. confidence judgements) and implicitly. As such, we recruited children between the ages of 4-8 years to explore what appears to be a period of significant metacognitive development.

Previous studies have not investigated markers of certainty in children aged 4-8 years after encoding a complex episodic event. Across two experiments, we tested whether and to what extent explicit and implicit metacognitive measures predict memory accuracy in children of different ages after children encoded a complex episodic event (similar to an eyewitness experience).

We had three main research questions: (1) which implicit measures were predictive of accuracy in children aged 4-8 years, (2) did any of these measures predict accuracy better than age, and (3) did any of these measures change in informativeness with age. We identified implicit measures to be predictive of accuracy from previous literature. Experiment 1 examines whether children of different ages can assign confidence that reflects their likely memory accuracy, and which other metacognitive measures are predictive of accuracy. We also investigated if the informativeness of these measures change with age. Experiment 2 furthers the results from Experiment 1: we explore if box-sorting decisions are more informative of accuracy when they are preceded with a confidence judgement in children of different ages. Ultimately, this research has important implications, such as informing legal decision makers how to better interpret children's memory evidence using metacognition.

Recent studies have shown that children engage in behaviours, such as gestures and vocalisations, that indicate uncertainty (e.g., Whitebread, et al., 2009), alongside response time and answer withholding. From this evidence, we can infer that younger child are sensitive to their uncertainty, and able to signal their uncertainty implicitly. If implicit measures predict memory accuracy for a complex event, then we would expect children to exhibit more of the certainty body and vocal measures, and fewer of the uncertainty body and vocal measures when they are accurate. We would also expect children to respond quicker and sort their answers into boxes when accurate. If this is the case, then it suggests that children can monitor their uncertainty and indicate their epistemic stance through these implicit behaviours. If such measures do predict accuracy, then it could be useful to implement these findings in a forensic setting. For example, if a head nod gesture significantly predicted accuracy, legal decision makers could be advised to note down when a child does this gesture at interview.

CHAPTER 2

EXPERIMENT 1

2.1 Introduction

As previously discussed, age is often used as a metric to determine memory accuracy, with young children commonly assumed to have less accurate memories than older children and adults (e.g., Knutsson & Allwood, 2014; Melinder, Goodman, Eilertsen, & Magnussen, 2004; Newcombe & Bransgrove, 2007; Wigmore, 1935). Young children are also believed to have poor *metacognitive abilities*, meaning that they are unable to monitor their uncertainty and distinguish when their memories are accurate and inaccurate (e.g., Keast, Brewer, & Wells, 2007; Powell, Garry, & Brewer, 2013). Together, the inference is that children's memories can often be inaccurate, and children are usually unaware of these inaccuracies. However, evidence from the developmental literature that suggests metacognitive abilities are present in children as early as the second year of life (Geurten & Bstin, 2018). In other words, children from perhaps the age of 2 years may be aware of when their memory is accurate and inaccurate (e.g., Balcomb & Gerken, 2008) and can indicate their uncertainty implicitly, for example by looking to their caregiver (e.g., Goupil, Romand-Monnier, & Kouider, 2016) or by shrugging their shoulders (e.g., Gelman & Bloom, 2000).

Good metacognitive monitoring means an individual can appropriately modulate their confidence in response to their performance (Fleming & Lau, 2014). If confidence is associated with memory accuracy (with high confidence indicating high accuracy, and lower confidence indicating lower accuracy), then confidence can be used as a predictor of memory accuracy. An outstanding question is whether children's metacognition—such as their confidence judgements—are predictive of memory accuracy.

The witness literature has typically concluded that children's confidence judgements do not reflect their memory accuracy (e.g., Allwood, Granhag, & Jonsson, 2006; Brewer & Day, 2005). Despite this, more recent eyewitness research and developmental research typically paints a more positive picture of the confidence-accuracy relationship in children, with children aged 7 years being able to accurately rate their confidence on eyewitness tasks (e.g., Bruer, Fitzgerald, Price, & Sauer, 2017; Winsor et al., in press), and children from age 4-5 years being able to express their confidence on simple memory tasks (e.g., Hembacher & Ghetti, 2014).

Children may have difficulty expressing their uncertainty explicitly using confidence for eyewitness tasks because the tasks are more difficult. Much of the relevant general developmental research has been conducted using simple decision-making tasks, and developmental memory research has used relatively simple encoding and test designs and materials, such as list-learning memory studies (e.g., Hembacher & Ghetti, 2014). Recent studies have shown that younger children engage in behaviours, such as gestures and vocalisations, that indicate uncertainty implicitly (see Harris, Bartz, & Rowe, 2017 for a review; Fusaro, Harris, & Pan, 2011; Hübscher, Vincze, & Prieto, 2019; Swerts & Krahmer, 2005; Visser, Krahmer, & Swerts, 2014).

From the evidence presented it appears that young children may be able to monitor the likely accuracy of their memories (Goupil, Romand-Monnier, & Kouider, 2016; (Monosov & Hikosaka, 2013), but have difficulty reliably indicating their epistemic stance explicitly, such as using confidence, on complex tasks. If children from age 4 years have good metacognitive awareness and uncertainty monitoring but are not yet able to use confidence scales as efficiently as 8-year-olds on complex memory tasks, then using implicit measures rather than explicit measures may provide more information on uncertainty (and therefore possibly memory accuracy) in younger children.

What is currently unknown is whether implicit measures of metacognition, such as body and vocal gestures, response time and answer exclusion, can be used to predict children's memory accuracy after children have encoded a complex episodic event. Some recent research suggests that this may be the case: for example, Winsor et al. (in press) found that children can indicate their uncertainty and certainty through use of interactive viewing behaviours on a lineup task (i.e., how children rotated the faces) after they had encoded either a video of a man tidying up toys or a mean returning home with shopping and eating chocolate. Yet, no previous research has examined the relative informativeness of a range of the implicit measures, nor compared them to informativeness of explicit measures (like confidence) in predicting accuracy after encoding a complex episodic task.

2.2 Study Aim

The aim of Experiment 1 was to examine which explicit (i.e., confidence) and implicit metacognitive measures (e.g., head nods) are predictive of memory accuracy on a complex episodic memory task in children aged 4-8 years and examine if the informativeness of the measures change with age. To do this, we identified implicit measures from previous developmental literature. Although not all literature focuses on the 4-8 age range, we use the evidence from the studies to support our hypothesis that these measures will indicate (un)certainty in 4-8-year-olds.

2.3 Implicit Measures of Metacognition

To determine which measures of implicit metacognition to examine in our episodic memory study, we reviewed the broader developmental literature and collated implicit measures of children's metacognition that had been identified in previous work. We collated 11 implicit measures in total, 9 of which were either body or verbal measures and we made a distinction between measures associated with (un)certainty. Two final implicit measure were the box-sorting task (Hembacher & Ghetti, 2014) and response time.

2.3.1 Certainty body measures

2.3.1.2 Head nods. Head nods are considered to signal high certainty (Roseano, González, Borràs-Comes, & Prieto, 2014; Vincze & Poggi, 2016), and the most stable gesture of certainty when compared to other certainty gestures (Borràs-Comes, Roseano, Bosch, Chen, & Prieto, 2011). Indeed, Harris, Bartz, & Rowe (2017) found that children in the second year of life associated head nods from adults with the correct location of an object, indicating they understand the gesture in the context of certainty. Children aged 3-5 also recognise head nods as signals of certainty in others (Hübscher, Esteve-Gilbert, & Igualada, 2017). The use of head nods has been observed early in childhood: Fusaro, Harris, & Pan (2011) coded for head gestures in children aged 14, 20 and 32 months during a semi-structure play session with their mother. Children in all three age groups used head nods to reinforce affirmative statements, but the frequency of usage was highest in 32-month-olds. Armstrong (2020) noted the use of head nods to indicate certainty in a 4-year-old child. Although there is little evidence for head nods as gestures of (un)certainty in the 4-8 age range, we can hypothesis that children within the 4-8 age range continue to use head nods as gestures of certainty.

2.3.2 Certainty vocal measures

2.3.2.1 Boosters. Boosters, such as ‘obviously’, ‘definitely’, ‘clearly’, are verbal markers of certainty that indicate a high commitment statement and individuals may use them to assert confidence (Hyland, 1998). Moore, Bryant, & Furrow, (1989) found in their study with children aged 3-8 years that all age groups understood the distinction between ‘know’ (a booster word) and ‘guess’ (a hedge word) and were aware that the former was indicative of certainty. As with head nods, children aged 3-5 also recognise boosters as signals of certainty in others (Hübscher, Esteve-Gilbert, & Igualada, 2017).From this, we could infer children appear to understand that booster words, such as ‘know’, are associated with certainty.

2.3.3 Uncertainty body measures

2.3.3.1 Shrugs. A shrug is a gesture that can be defined by multiple movements: moving the shoulders upwards, rotating the forearms upwards, and flipping hands over (‘palm up open hand gesture’). Shrugs can include an isolated feature or a combination the features (Kim, Paulus, Sodian, & Proust, 2016). A shrug has been described to be a densely communicative behaviour that can convey various meanings (Debras, 2017; Givens, 1977), including ignorance and uncertainty (Poggi, 2016). Shrugs were often accompanied with the verbal marker ‘I don’t know’ in 66 children aged 3-5 years during an object description task (Gelman & Bloom, 2000). Further, in clinical observations of 65 child patients, children were found to shrug when they did not want to participate in discussion, mirroring withholding behaviours when uncertain (Wassmer, et al., 2004).

2.3.3.2 Looking to caregiver. When a child looks to caregiver, it is assumed they are asking for help or looking to see if the adult has useful information (Campos & Steinberg, Perception, appraisal and emotion: The onset of social referencing., 1981). Various studies have demonstrated that children seek help by making eye contact with their caregiver (e.g., Kim & Kwak, 2011; Poulin-Dubois & Brosseau-Liard, 2016). In a study by Goupil & colleagues (2016) 80 children aged 19-21 months looked towards an adult to 'ask for help' during a non-verbal memory monitoring paradigm involving locating a hidden toy. Indeed, it is believed that a child's metacognitive abilities are facilitated by their primary caregiver, as children internalise the cognitive guidance provided by them (Smith, Shields, & Washburn, 2003).

2.3.3.3 Thinking gesture. A thinking gesture has been defined as a closed hand and index finger touching, tapping or leaning on the lower half of the face (Mahmoud & Robinson, 2011). Mahmoud and Robinson (2011) coded videos depicting spontaneous facial expressions of 12 adults during various tasks, and they identified that a thinking gesture was associated with uncertainty. This gesture was observed in the pilot data, and so we included it on an exploratory basis. Additionally, Hembacher & Ghetti (2014) used a confidence scale that depicted a cartoon of a child displaying a thinking getsure at the low end of the scale. Children from the age of 4 years seemed able to use the scale appropriately, suggesting that they recognise the gesture as a indicator of uncertainty.

2.3.3.4 Head shake. A head shake involves moving one's head from side to side. The movement can be small or exaggerated and can involve one 'shake' or multiple shakes. Research suggests that children begin to use head gestures between 8 and 14 months (e.g., Goldin-Meadow, 2015), with head shakes preceeding head nods (Fusaro & Harris, 2013). Fusaro & Harris (2013) found that 24-month-olds could differentiate between head nods and shakes in adults and the inferences that came with these gestures. Fusaro, Harris, & Pan (2011) also found that children as young as 14 months performed headshakes when they did not know the answer to their mother's question.

2.3.3.5 Head tilt. A head tilt involves an individual tilting their to the side, backwards or forwards and has been identified as being associated with uncertainty (Borràs-Comes, Roseano, Bosch, Chen, & Prieto, 2011). Kim (2016) tested the sensitivity of 36 3- and 4-year-olds to their own ignorance through an informing task where children were asked to help an adult identify an item placed in a box. When children were uncertain of which item had been placed in the box, they often tilted their head to the side (Kim, 2016). As with head nods, children aged 3-5 also recognise head tilts as signals of uncertainty in others (Hübscher, Esteve-Gilbert, & Igualada, 2017).

2.3.4 Uncertainty vocal measures

2.3.4.1 Hedges. Hedges are statements that convey low commitment and express uncertainty about a proposition (Gustafsson, Lindholm, & Jönsson, 2019; Holmes, 1990; Lakoff, 1975), such as 'I think', 'could be', 'maybe', 'I don't know'. Others have described hedges being used in anticipation of an overstatement. In other words, hedges may act as a buffer to a statement that the individual has assessed to be potentially to be incorrect. Peterson & Briggs (2001) interviewed children aged 3-, 5-, 8-years-old about a time they felt a particular emotion and hedges were identified as a marker of uncertainty, with the frequency of usage increasing with age. As with head nods and head tiltd. children aged 3-5 also recognise hedges as signals of uncertainty in others (Hübscher, Esteve-Gilbert, & Igualada, 2017).

2.3.4.2 Fillers. Fillers are non-word utterance expressed without clear meaning, such as ‘um’, ‘uh’, ‘erm’, ‘hmm’. Fillers are often used to fill a silence. Fillers are often observed in adult speech as markers of uncertainty (Clark & Tree, 2002). Krahmer & Swerts, (2005) found in his study of adults and 8-year-old children that, whilst fillers were a more robust marker of uncertainty in the adults, children still produced fillers when uncertain, and could also detect in them others.

2.3.5 Other implicit measures

2.3.5.1 Response time. There is ample previous literature that suggests response time is indicative of certainty and uncertainty. Patterson, Cosgrove, & O'Brien, (1980) used messages of varying complexity to explore 6-, 8- and 10-year-old’s nonverbal gestures of uncertainty. The results indicated that children of all ages had longer response times when the message was complex. Roderer & Roebers (2010) tested 7- and 9-year-olds uncertainty monitoring using a vocabulary retrieval task and found that children looked for longer at items that were more difficult, indicating a longer response time when unsure about their knowledge. Leckey, et al., (2020) came to a similar conclusion with 25- to 32-month-olds, who looked longer at items that they were unsure of, indicating children hesitate when they encounter uncertainty (Beck, Robinson, & Freeth, 2008). For response time, we measured the time the questioned appeared on the screen to when the child moved their arm to point at the screen.

2.3.5.2 Box sorting. Children have consistently been shown to withhold or withdraw answers or skips trials when unsure of their answer (e.g., Lyons & Ghetti, 2013; Kim, 2016). For example, Krebs & Roebers, (2010) and Koriat and collegaues (2001) tested children between the ages of 7-12 years and both concluded that children could successfully identify and withhold inaccurate answers, indicating the increasing developmental trajectory of uncertainty monitoring. Specifically, a box sorting task has demonstrated how children choose to withhold their low confidence answers **(**Hembacher & Ghetti, 2014).

2.4 Summary

If implicit measures predict memory accuracy for a complex event, then we would expect children to exhibit more of the certainty body and vocal measures, and fewer of the uncertainty body and vocal measures when they are accurate, and vice versa. We would also expect children to respond quicker and sort their answers into the open-eye box when accurate. If this is the case, then it suggests that children can monitor their uncertainty and express their epistemic stance through these implicit behaviours. If such measures do predict accuracy in a forced choice task after a complex episodic event, then it could be useful to further investigate these measures using techniques utilised by legal professionals to interview child witnesses (e.g., free-recall or cued-recall) to see if they could be collected in a legal setting. For example, if a head nod gesture significantly predicted accuracy, legal decision makers could be advised to note down when a child does this gesture at interview.

2.5 Experiment 1: Method

2.5.1 Design

We used a within-subject design. Video order was counterbalanced. Question order was randomised for each participant. Our data-collection stopping rule was to recruit at least 40 children. The research was reviewed according to the University of Birmingham Science, Technology, Engineering and Mathematics Ethical Review Committee.

2.5.2 Participants

A total of 50 neurotypical children between the ages of 4-8 years were recruited via advertising on social media platforms (Facebook, Twitter, Linkdin) and recruitment websites (childrenhelpingscience.com, callforparticipants.com, honeybee.io, psych.hanover.edu). Data from 10 participants were excluded, resulting in a final sample size of 40. Data was excluded due to technical difficulties (e.g., poor internet connection, poor video quality), due to neurodevelopmental conditions that were subsequently reported by guardians/parents, and due to not understanding the instructions for the cup scale. This resulted in a sample of 40 participants (*M* age = 5.93, *SD* age = 1.42; 50% female). Each experimental session had a maximum of 40 trials. Out of the 40 included participants, nine children chose to stop participating at various points during the experiment, resulting in 165 incomplete trials. An additional five trials were excluded from three children due to adult interference. Therefore, a total of 170 trials were removed from the dataset of 40 participants, leaving 1,430 completed trials to be analysed. Children received a £10 Amazon gift card and a certificate as a reward for taking part.

2.5.3 Materials

We filmed two videos for the encoding phase in the study. Each video was approximately one minute long and depicted either an adult making breakfast or an adult washing up. For the test phase, we created a 2AFC, which involved a question with two responses to choose from. Leaving a question unanswered was not possible. For each video event there were 20 questions (see Appendix C and Appendix D). For each question, we created two images as response options: an image of the item that was in the video and an alternative image of an item that was not. For the items that were in the video, we photographed the items exactly as they were presented in the videos (e.g., the front of the fridge, as it was presented in the video). The alternate images not seen in the video were photographed separately, but from the same angle as those present in the video (e.g., a different fridge but also shown from the front). All images were formatted in the same way (pixel quality, rotation of image, lighting) so that the someone who had not seen the video could not discriminate between items that were versus were not in the video. The 20 questions for each video were created to range in difficulty; for example, some questions were about central items in the video (e.g., which top was the boy wearing?), whereas others were about background items in the video (e.g., what else was in the fridge?). Pilot testing on three children aged 4-, 5-, 8- years, showed that average accuracy was above chance level (50%) for all children (mean % correct = 80%, with children scoring 85% correct, 73% correct, and 83% correct, respectively). This pilot testing satisfied us that the questions were appropriate for 4–8-year-olds.

To measure confidence, we used the 5-point Likert water cup scale from Bruer et al (2017) (adapted from the Cup Scale: Weston, Boxer, & Heatherington, 1998). The water in the cups increased across the scale to represent increasing confidence. Unlike Bruer's scale, there were no numbers on the scale, because the older children might be more proficient at reading numbers than the younger children in our sample. Research has indicated that children find pictorial scales easier to use than numerical scales (e.g., Ghetti, Hembacher, & Coughlin, 2013). Children were told that the cups would measure how sure they were that their answer was right, and that the surer they were in their decision, the more water in the cup. Each level of certainty was explained carefully; children were told that if they were not sure in their decision, they should choose a cup that did not have very much water or any water in it. If they were a little bit sure but not too sure in their decision, they should choose a cup that had some water in it, but not totally full. If they were very sure in their decision, they should choose a cup that is almost or totally full.

We used the box sorting task detailed in Hembacher & Ghetti (2014): one box had an open eye on it, and the other box had a closed eye on it. Children were told that they would also have to choose whether to put their answer into one of two boxes (an open-eye box or a closed-eye box). Children were told that the answers in the closed-eye box would not be checked later, and that if they did not want us to look at their answer, to put it in the closed-eye box, and that answers in the open-eye box would be checked later, and that if they did want us to look at their answer, to put it in the open-eye box. Likewise, we said it would be a good idea to put correct answers in this box.

2.5.4 Procedure

Informed consent was given by guardians. One experimenter tested all children to ensure consistency across participants. Testing session were recorded on Zoom. Participants were required to use a computer or a laptop for a consistent webcam set up across participants and resulted in clear recordings for subsequent coding. We asked the adult to share their screen so their progress could be followed. A child assent form was completed at the beginning of the testing session (e.g., children were asked “do you have any questions you’d like to ask” and “have all your questions been answered in a way you understand?”). If the child had additional questions, the experimenter answered these before proceeding. After obtaining assent, we collected the child’s age, gender and first language. We also assigned them an ID number should the parent want to later withdraw their data.

To set up the webcam, we provided instructions and images via screen share to ensure that the adult and child were sat so they could both be seen clearly in the recordings. We instructed the child to sit in the centre of the screen with their torso in shot, so all upper body gestures could be seen for coding. We also instructed that the adult remain in the shot if possible, so that we could see if they helped or guided the child at any point, and subsequently disregard that trial in the analysis.

The children were told they were going to play a game that involved watching two short videos of simple daily tasks. They were asked to watch the videos carefully, as they would be asked questions about what they could remember afterwards. We showed participants two example images to give them an idea of what the stimuli would be like in the question phase. The children were told that they would have rate how sure they were that their answers were right and introduced the confidence scale. We then asked what their name was and asked them to rate how sure they were that they were right using the cups. Only when the child seemed to understand the scale did the experimenter move on.

Next, we told the children that they would also have to choose whether to put their answer into one of two boxes (an open-eye box or a closed-eye box). Children were told that to get a good prize they had to get a lot of correct answers, but not many incorrect answers. We explained it would be a good idea to put incorrect answers in this box, and then explained that answers in the open-eye box would be checked later, and that if they did want us to look at their answer, to put it in the open-eye box. Likewise, we said it would be a good idea to put correct answers in this box. Children were told that their prize would depend on their correct answers in the open-eye box. Children were asked to point at their answers, and the adult was asked to control the mouse and click on the answer that the child had indicated. Children of different ages might be differentially proficient at using a computer mouse, but presumably equally able to 'point' at their answer. Given that children's response time was a variable of interest, we asked adults to control the mouse to account for mouse use bias across age groups. Adults were asked to refrain from helping their child. At the start of each question a blue 'home' symbol appeared at the top of the screen. Adults were asked to return and keep the cursor here until the child had chosen their answer, to prevent the child being guided towards a response by the adult's mouse movement and positioning (see Appendix A for experiment set up).

Before starting the Experiment, children had a practice round which involved watching a 6 second practice video and answering 4 practice questions. Each practice question required a 2AFC response, a confidence rating on the water-cup scale, and a box-sorting decision. In the practice round, each decision was narrated after it was made: if the child chose the cup with no water in it, the experimenter would say, for example, "so you're not sure at all of your decision". We gave feedback for each stage of every question. If the child used the scales correctly, they were told well done and that they were good at using the scales. If it seemed they did not understand the scales and used them incorrectly, we would remind them what the scales meant. These instructions were repeated until the child demonstrated a good understanding of how to use the scales.

After the practice round, each child watched one of the two videos. The order of the two videos was counterbalanced across participants. After watching the video, we asked the child if they remembered what they had to do, then started the first question phase of 20 questions. In the question phase, we read aloud each question in a neutral voice. We read the questions aloud to avoid any reading bias (i.e., older children being able to read the question more efficiently, and so answering quicker) that may have occurred if the questions appeared on screen. In the question phase, question order, the position of the two answers (left vs. right), and the position of the open-eye and closed-eye box (left vs. right) were randomised for each trial. The confidence scale remained in the same position throughout the Experiment, with the empty cup displayed at the left. The boxes were randomised to discourage children from choosing the same box for every question, or for simply choosing a box directly under the cup they had chosen on the confidence scale. We also hoped that randomising the box order would keep the children engaged throughout the question phase; rather than anticipating where the box would be and pointing to the same place on the screen each time, children would have to check which order the boxes appeared in. Children were encouraged every few questions (e.g., “you’re doing great”). From our experience in the pilot experiment, this kept the children focused on the task, and ensured that the children’s answers were not guided by what the experimenter was asking. It was also hoped to help motivate the child to finish the task, as we also let them know they had nearly finished (e.g., “well done, you’re nearly finished now”) when they were towards the end of the task.

After the first question phase was complete, children were given the opportunity to take a break to reduce fatigue, or to stop if they did not want to continue. After the break, they watched the second video, and answered a second set of 20 questions. Therefore, each child completed a maximum of 40 trials. After completion, we congratulated the child on finishing, asked if there were any questions, and provided the debrief information.

2.5.5 Results

The aim of our study was to examine which explicit (i.e., confidence) and implicit (i.e., vocal and body gestures) metacognitive measures are predictive of memory accuracy on a complex episodic memory task and examine if the informativeness of the measures change with age. This section will first provide an overview of memory performance on the task and then describe the implicit metacognitive measures and how they were coded. We Z-transformed each measure and plotted them in a chart to visualise the frequency of measures observed when accompanying correct and incorrect answers. Then we describe a series of multilevel logistic regressions which were run to see which measures significantly predictive accuracy, and if the informativeness of these measures changed with age. Finally, we plotted accuracy characteristic plots for the statistically significant measures from the regression to visualise the relationship between the measure and accuracy in younger and older children.

2.5.5.1 Memory Accuracy

There were 25 younger children (aged 4-6 years, *M* age = 4.96, SD = .79) and 15 older children (aged 7-8 years, *M* age = 7.53, *SD* = .52). Performance on the memory task was high for all ages (*M* = .80, *SD* = 40). Younger children's mean accuracy (*M* = .77, *SD* = 0.42) was -.06 (*SE* = .03) lower than older children's (*M* = .83, *SD* = .37). An independent-samples t-test was run to determine if there were significant differences in accuracy between younger and older children. The difference in mean accuracy between age groups just reached statistical significance, *t*(38) = 2.12, *p* = .04.

2.5.5.2 Coding

A coding protocol detailing the 9 vocal and body implicit metacognitive measures was designed for the study (see Table 1 for a description of each measure). We also coded response time as the time between the question appearing on screen and the child pointing to their answer and the final implicit measure was the box-sorting decision which was recorded by Qualtrics and did not require coding (total = 11 implicit measures). Measure frequency was coded 0 as ‘not present’ and =>1 as ‘present’ and entered a series of multi-level logistic regressions. As an explicit measure we collected confidence which was also recorded by Qualtrics.

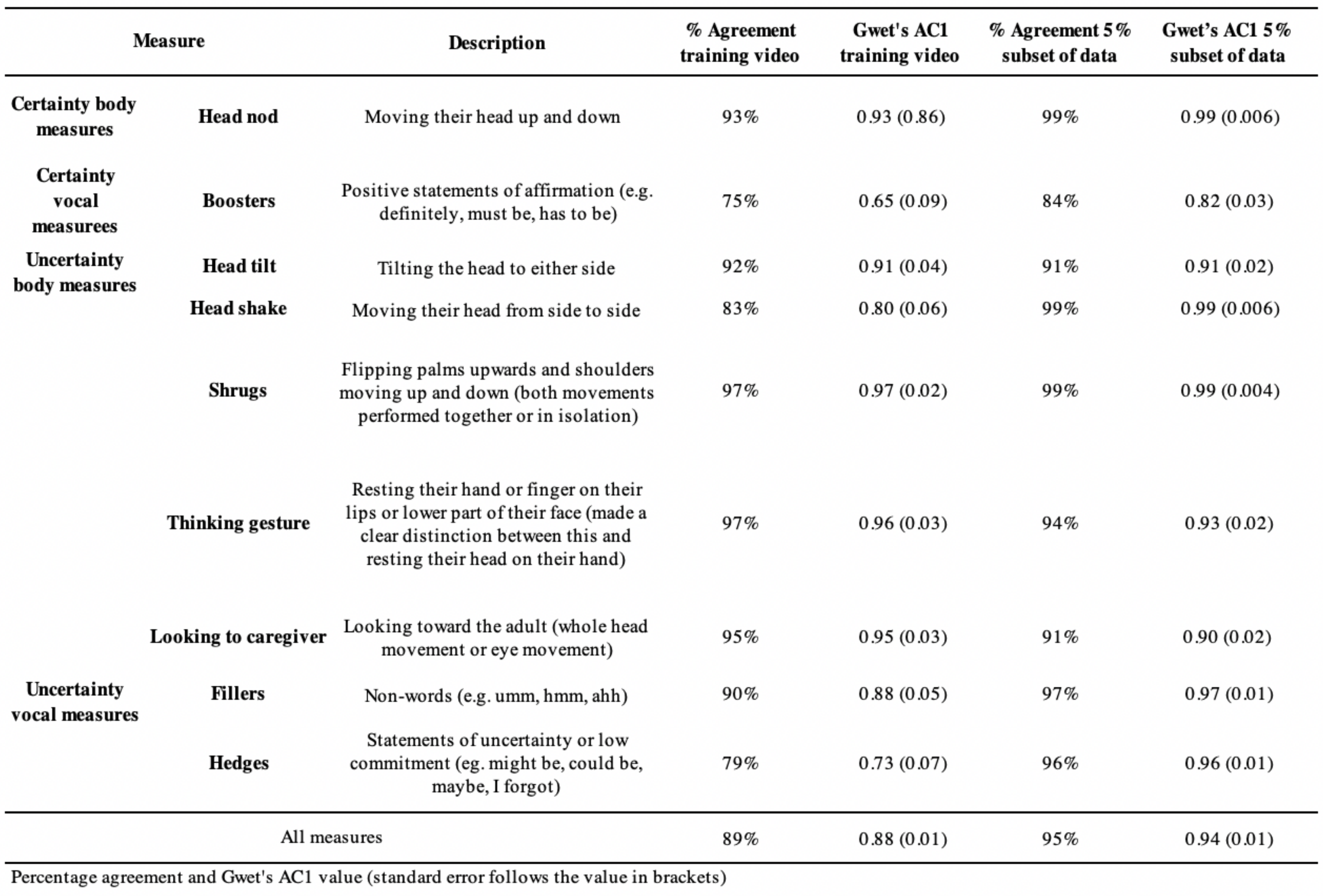
Coders blind to the purpose of the study were recruited to code implicit measures displayed in the videos of the child participants. The protocol was explained to the coders; each measure was introduced and discussed in detail. They were asked to code for behaviours that occurred when the child was answering the 2AFC memory question. Behaviours were not coded when the child was answering the confidence or box sorting questions. Each coder then individually coded the testing session of one participant who completed 40 trials. There was a high percentage of agreement between coders (89% overall).

We wanted to quantify the interrater reliability on our coding scheme. Research has shown that Kappa skews data when there is a high agreement between coders (known as the Kappa Paradox (P_0); Feinstein & Cicchetti, 1989; Gwet, 2008). To avoid the Kappa Paradox, we used Gwet’s AC_1 coefficient to calculate the interrater reliability, which is more robust to high agreement among raters (Gwet, 2001; Gwet, 2002). For the first video coded, the overall interrater reliability across the three coders was AC_1= .88. For Gwets AC_1, the benchmarks are .80 – 1.00 for very good, .60 – .80 for good, .04 – .60 for moderate, .20 – .40 for fair, and

< .02 for poor (Gwet, 2014). The Gwets AC_1 results for each implicit measure coded in the training video is shown in Table 1.

Once training was complete, each coder independently coded the data for approximately 13 participants. In addition, all coders coded the same 5% of participants (*N* = 5 participants) and we again measured the interrater reliability, which was AC_1 = .94 overall. The interrater reliability for all individual measures were also calculated and are presented in Table 1. In short, these results illustrate that the coders agreed about the measures that they coded.

Table 1. The percentage of agreement and Gwet's AC1 reliability score of all coded measures for training video and 5% subset of the data



Measure		Description	% Agreement training video	Gwet's AC1 training video	% Agreement 5% subset of data	Gwet's AC1 5% subset of data
Certainty body measures	**Head nod**	Moving their head up and down	93%	0.93 (0.86)	99%	0.99 (0.006)
Certainty vocal measurees	**Boosters**	Positive statements of affirmation (e.g. definitely, must be, has to be)	75%	0.65 (0.09)	84%	0.82 (0.03)
Uncertainty body measures	**Head tilt**	Tilting the head to either side	92%	0.91 (0.04)	91%	0.91 (0.02)
	Head shake	Moving their head from side to side	83%	0.80 (0.06)	99%	0.99 (0.006)
	Shrugs	Flipping palms upwards and shoulders moving up and down (both movements performed together or in isolation)	97%	0.97 (0.02)	99%	0.99 (0.004)
	Thinking gesture	Resting their hand or finger on their lips or lower part of their face (made a clear distinction between this and resting their head on their hand)	97%	0.96 (0.03)	94%	0.93 (0.02)
	Looking to caregiver	Looking toward the adult (whole head movement or eye movement)	95%	0.95 (0.03)	91%	0.90 (0.02)
Uncertainty vocal measures	**Fillers**	Non-words (e.g. umm, hmm, ahh)	90%	0.88 (0.05)	97%	0.97 (0.01)
	Hedges	Statements of uncertainty or low commitment (eg. might be, could be, maybe, I forgot)	79%	0.73 (0.07)	96%	0.96 (0.01)
All measures			89%	0.88 (0.01)	95%	0.94 (0.01)

Percentage agreement and Gwet's AC1 value (standard error follows the value in brackets)

2.5.5.3 Z-transformed measures

First, to visualise our data, we first Z-transformed the mean amount of each measure and plotted them as a function of accuracy, following Gustafsson, Lindholm, & Jönsson, (2019). Figure 2 shows that, descriptively speaking, when children were incorrect, they made lower confidence judgements, had increased response times, and more frequently chose to hide their answers in the box-sorting task. Children also performed more head tilts, thinking gestures, hedged more and used more fillers when their answers were incorrect. Interestingly, and in contrast to the predicted direction, head nods seemed to be performed more when children were incorrect rather than correct. Children did not perform many head shakes, shrugs or looking to caregiver gestures when incorrect. When children were correct, they made higher confidence judgements, and had faster response times. They also performed more boosters but did not perform many head nods.

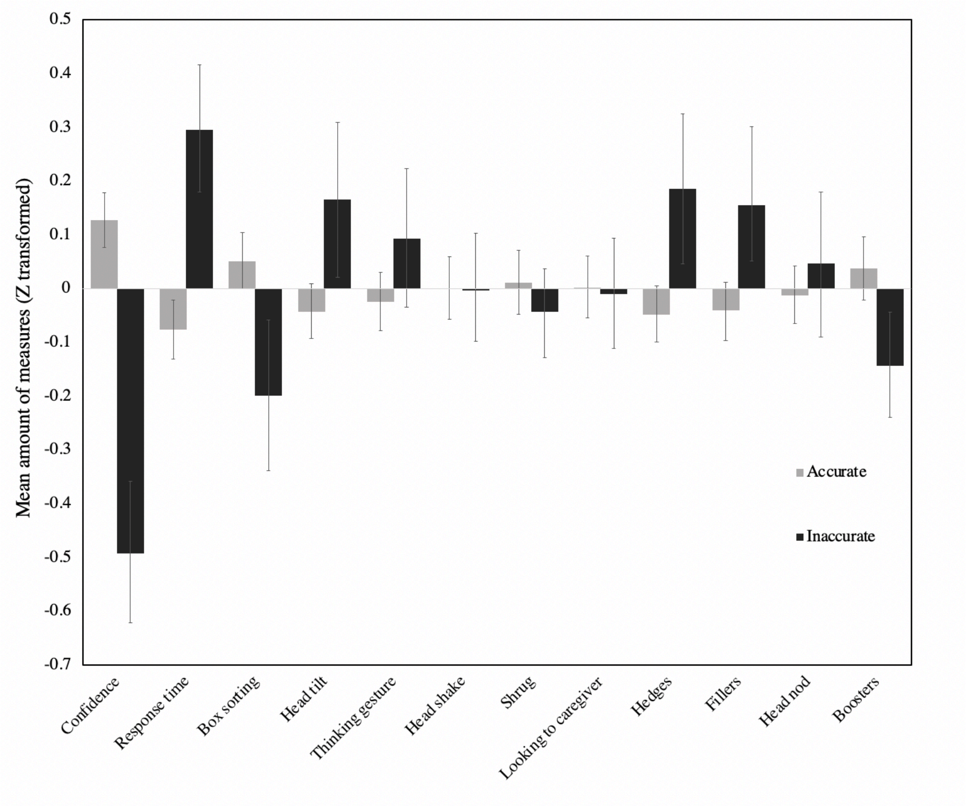


Figure 2. Mean amount of implicit and explicit measures (z-transformed) for accurate and inaccurate answers for Experiment 1. Error bars represent 95% confidence intervals

2.5.5.4 Multilevel modelling

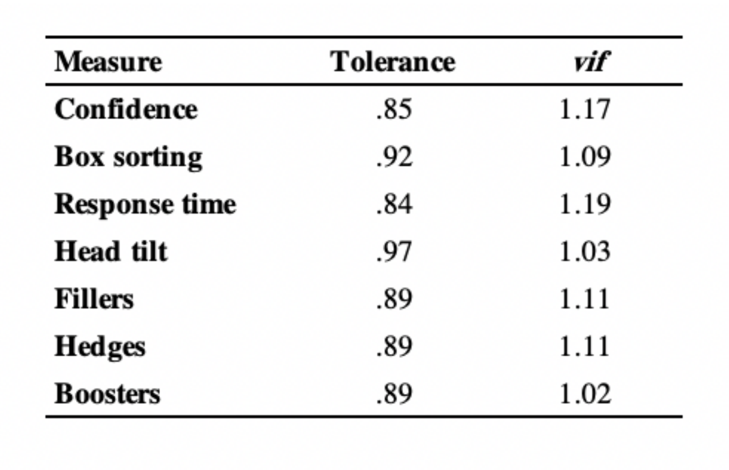
Next, to answer how predictive the measures were of accuracy, we fitted a series of multilevel logistic regression models using the lme4 package in R. We organised the data as multilevel data with individual responses nested within participants and followed the methods used in Gustafsson, Lindholm, & Jönsson (2019) and Mansour, Beaudry, & Lindsay (2017). To assess model fit, we used the likelihood ratio test as it is the most liberal test and is typically used in eyewitness research (e.g., Horry, Halford, Brewer, Milne, & Bull, 2014; Wright & London, 2010).

2.5.5.4.1 Examining which measures predict memory accuracy

To examine which of the measures were significantly predictive of accuracy, we compared models for each of the 12 measures (models 2-13) to a null intercept-only model predicting accuracy (model 1). Table 4 reports the model parameters and fit indices for models 1-14. Note that tests for skewness indicated that some of the data were not normally distributed.

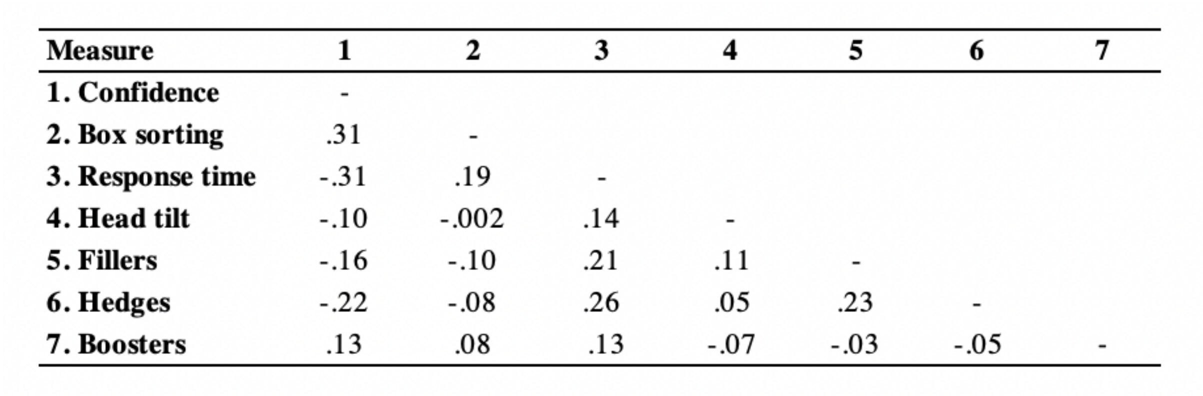
Compared to the null model (model 1, intercept), the model fit was significantly improved when adding confidence, $\chi^2(1) = 103.27$, $p < .001$ (model 2); box sorting, $\chi^2(1) = 15.75$, $p < .001$ (model 3); response time, $\chi^2(1) = 29.37$, $p < .001$ (model 4); head tilt, $\chi^{2(}1) = 10.65$, $p <.01$ (model 5); hedges, $\chi^2(1) = 13.47$, $p < .001$ (model 10); fillers, $\chi^2(1) = 9.21$, $p < .01$ (model 11); $p < .01$; and boosters, $\chi^2(1) = 8.72$, $p = .003$ (model 13), but not by head shakes, $\chi^2(1) = .01$, $p = .79$ (model 6); shrugs, $\chi^2(1) = .80$, $p = .37$ (model 7); thinking gestures, $\chi^2(1) = 3.49$, $p = .06$ (model 8); looking to caregiver, $\chi2(1) = .001$, $p = .97$ (model 9); or head nods, $\chi^2(1) = .79$, $p = .37$ (model 12). Tests to see if the data for the 7 significant predictors met the assumptions of collinearity indicated that multicollinearity was not a concern. Tolerance values below .2 and vif values above 4 indicate a problem (Menard, 1995; Hari, 2010). Tolerance and *vif* values for the 7 significant predictors indicated no incident of multicollinearity (see Table 2 & Table 3).

Table 2. Significant measures tolerance and *vif* statistics for Experiment 1



Measure	Tolerance	*vif*
Confidence	.85	1.17
Box sorting	.92	1.09
Response time	.84	1.19
Head tilt	.97	1.03
Fillers	.89	1.11
Hedges	.89	1.11
Boosters	.89	1.02

Table 3. Significant measures correlation matrix for Experiment 1



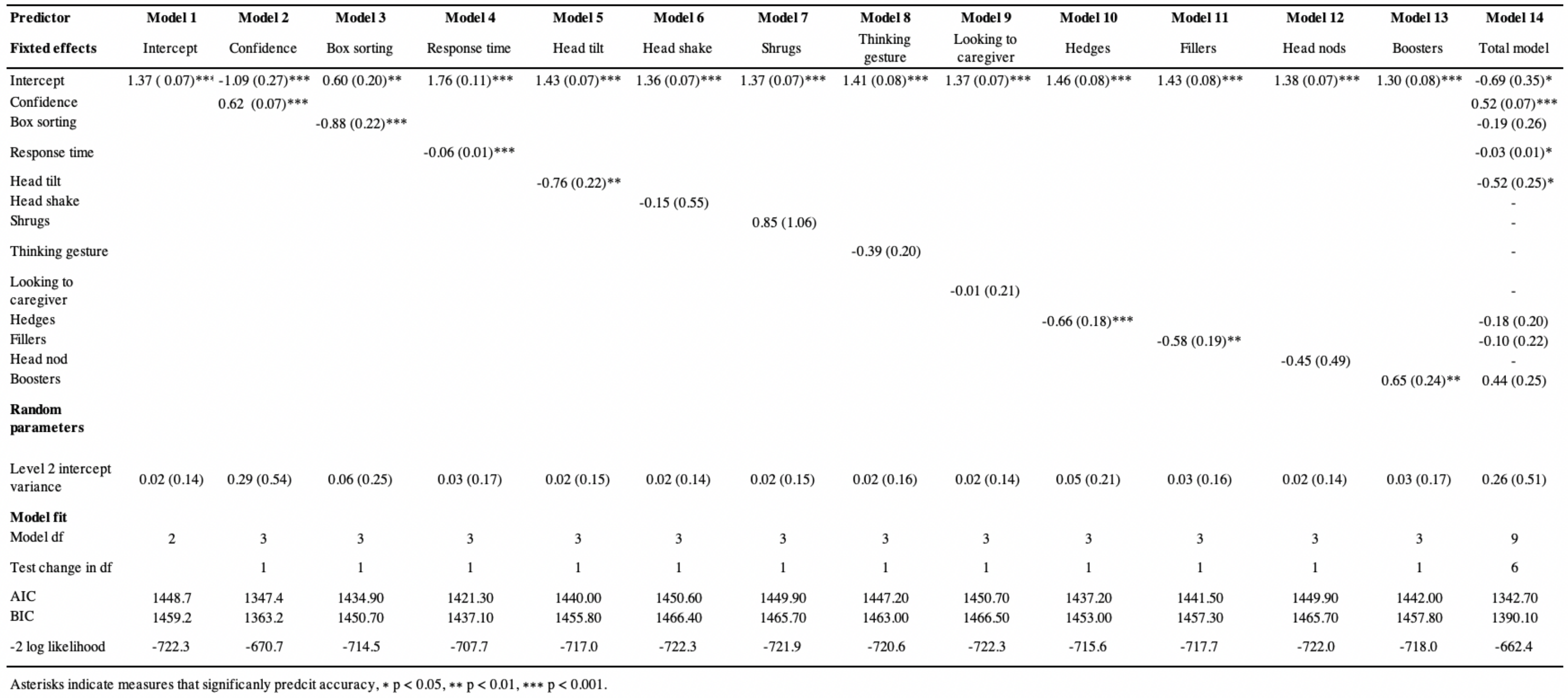
Measure	1	2	3	4	5	6	7
1. Confidence	-						
2. Box sorting	.31	-					
3. Response time	-.31	.19	-				
4. Head tilt	-.10	-.002	.14	-			
5. Fillers	-.16	-.10	.21	.11	-		
6. Hedges	-.22	-.08	.26	.05	.23	-	
7. Boosters	.13	.08	.13	-.07	-.03	-.05	-

The 7 significant predictors were added to one model (model 14), which was then compared to each significant predictor model separately. The model with all 7 significant predictors (model 14) improved the fit relative to the confidence model, $\chi^2(6) = 11.22$, $p =$ .08 (model 2); box sorting model, $\chi^2(6) = 98.73$, $p < .001$ (model 3); response time model, $\chi^2(6) = 85.11$, $p < .001$ (model 4); head tilt model $\chi^2(6) = 103.82$, $p < .001$ (model 5); hedges model, $\chi^2(6) = 101.01$, $p < .001$ (model 10); fillers model, $\chi^2(6) = 105.26$, $p < .001$ (model 11); and boosters model, $\chi^2(6) = 105.75$, $p < .001$ (model 13), indicating that a model with all significant predictors better predicted accuracy than each of the significant predictors alone.

2.5.5.4.2 Examining which uniquely measures predict memory accuracy

Next, we were interested in comparing the predictive performance across the measures. We z-transformed each of the significant measures and used the z-transformed values for all subsequent analyses. We first examined which measures uniquely predicted accuracy when controlling for the other predictors, so we fit the model with all 7 significant predictors (model 14), but this time used the z-transformed measures (model 15). Table 5 reports the model parameters and fit indices for models 15-24 using the z-transformed measures. Confidence (z = 7.41, p < .001), response time (z = -2.08, p = .04), and head tilts (z = -2.07, p = 0.04) uniquely explained memory accuracy when controlling for the other predictors (model 15). Specifically, as accuracy increased, confidence increased, and response time decreased. Box sorting (z = .73, p = .46); fillers (z = -.46, p = .65); boosters (z = 1.75, p = .08); and hedges (z = -.87, p = .38) did not uniquely explain memory accuracy when controlling for the other predictors. This suggests that, although a model with all predictors predicts accuracy better than any of the significant predictors alone, the unique predictors each contribute a percentage of the variance that can be captured only by them alone. That is, the collection of confidence, response time, and head tilts provides unique information to the prediction of accuracy over and above the other significant measures. Conversely, the collection of fillers, hedges, and boosters does not provide unique information can cannot be captured by the other measures.

Table 4. Model parameters of all measures when examining which measures predict accuracy for Experiment 1



Predictor	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14
Fixted effects	Intercept	Confidence	Box sorting	Response time	Head tilt	Head shake	Shrugs	Thinking gesture	Looking to caregiver	Hedges	Fillers	Head nods	Boosters	Total model
Intercept	1.37 (0.07)***	-1.09 (0.27)***	0.60 (0.20)**	1.76 (0.11)***	1.43 (0.07)***	1.36 (0.07)***	1.37 (0.07)***	1.41 (0.08)***	1.37 (0.07)***	1.46 (0.08)***	1.43 (0.08)***	1.38 (0.07)***	1.30 (0.08)***	-0.69 (0.35)*
Confidence		0.62 (0.07)***												0.52 (0.07)***
Box sorting			-0.88 (0.22)***											-0.19 (0.26)
Response time				-0.06 (0.01)***										-0.03 (0.01)*
Head tilt					-0.76 (0.22)**									-0.52 (0.25)*
Head shake						-0.15 (0.55)								-
Shrugs							0.85 (1.06)							-
Thinking gesture								-0.39 (0.20)						-
Looking to caregiver									-0.01 (0.21)					-
Hedges										-0.66 (0.18)***				-0.18 (0.20)
Fillers											-0.58 (0.19)**			-0.10 (0.22)
Head nod												-0.45 (0.49)		-
Boosters													0.65 (0.24)**	0.44 (0.25)
Random parameters														
Level 2 intercept variance	0.02 (0.14)	0.29 (0.54)	0.06 (0.25)	0.03 (0.17)	0.02 (0.15)	0.02 (0.14)	0.02 (0.15)	0.02 (0.16)	0.02 (0.14)	0.05 (0.21)	0.03 (0.16)	0.02 (0.14)	0.03 (0.17)	0.26 (0.51)
Model fit														
Model df	2	3	3	3	3	3	3	3	3	3	3	3	3	9
Test change in df		1	1	1	1	1	1	1	1	1	1	1	1	6
AIC	1448.7	1347.4	1434.90	1421.30	1440.00	1450.60	1449.90	1447.20	1450.70	1437.20	1441.50	1449.90	1442.00	1342.70
BIC	1459.2	1363.2	1450.70	1437.10	1455.80	1466.40	1465.70	1463.00	1466.50	1453.00	1457.30	1465.70	1457.80	1390.10
-2 log likelihood	-722.3	-670.7	-714.5	-707.7	-717.0	-722.3	-721.9	-720.6	-722.3	-715.6	-717.7	-722.0	-718.0	-662.4

Asterisks indicate measures that significanly predcit accuracy, * p < 0.05, ** p < 0.01, *** p < 0.001.

2.5.5.4.3 Examining which measures predict memory accuracy better than age

We examined if any measures were more predictive of accuracy than age. To this end, we took the z-transformed model with 7 predictors (model 15; Table 5) and added age as a main effect (model 16). Age was also z-transformed. Adding age as a main effect significantly improved the model fit, indicating that memory accuracy improved with age. We now refer to this model as the best-fitting main effects model. Confidence ($z = 7.77$, $p < .001$), head tilts ($z = -2.20$ $p = .03$) and age ($z = 4.09$, $p < .001$) all uniquely explained memory accuracy when controlling for the other predictors. As accuracy increased, confidence increased, head tilts decreased, and age increased. Notably, the size of the *z* scores indicate that *confidence was more informative of accuracy than age.* When age was added to the model, *response time was no longer a unique predictor of accuracy*, suggesting that age accounts for the portion of variance that was previously uniquely represented by response time.

2.5.5.4.4 Examining if the informativeness of the measures change with age

To examine if the informativeness of the measures change with age, we investigated if adding the fixed interaction of age separately with each significant predictor (model 17-23) would improve the model fit compared to the best-fitting main effects model (model 16). Only adding the interaction between age and confidence, χ^2 (11) = 9.08, $p = .003$ (model 17); the interaction between age and box sorting, $\chi^2(11) = 3.98$ $p = .05$ (model 18); and the interaction between age and head tilt, $\chi^2(11) = 4.62$, $p = .03$ (model 20) improved the model fit. This suggests that the informativeness of confidence, head tilt and box sorting in predicting memory accuracy changes with age. Adding the interaction between age and response time, $\chi^2(11) = .001$, $p = .99$ (model 19); age and hedges, $\chi^2(11) = .05$, $p = .83$ (model 21); age and fillers, $\chi2(11) = .17$, $p = .68$ (model 22); and age and boosters, $\chi^2(11) = 0.81$, $p = .37$ (model 23) did not improve the model fit compared to the best-fitting main effects model. This suggest that the informativeness of response time, boosters, hedges, and fillers in predicting memory accuracy do not change significantly with age.

Finally, to examine if a model with multiple interactions with age (i.e., confidence, head tilt and box sorting) was better than models with one interaction, we took the best-fitting main effects model (model 16) and added interactions between both age and confidence, age and box sorting, and age and head tilts (model 24). The model with all three interactions was better than the model with only the interaction between age and confidence, $\chi^2(13) = 5.32$, $p = .07$ (model 17); the model with only the interaction between age and box sorting, $\chi^2(12) = 10.41$, $p = .005$ (model 18); and the model with only the interaction between age and head tilts, $\chi^2(12) = 9.78$, $p = .01$ (model 20). Therefore, our final model—the best fitting interaction model—included confidence, head tilt, box sorting, fillers, hedges, boosters, response time and age as main effects, and the interactions between age and confidence, age and head tilts, and age and box sorting.

To interpret the three significant interaction effects, we initially planned to run a series of multi-level logistic regressions identical to the first series, for the younger (aged 4-6 years) and older (aged 7-8 years) child age groups separately. The sample size from Experiment 1 alone, however, was not deemed to be large enough to do this because the models did not converge. In short, we were trying to estimate too many coefficients with too few observations and overfitted models may have poor power (Bates et al., 2015; Dale et al., 2013). Instead, here, we visualise the interaction findings using accuracy characteristic plots.

Table 5. Model parameters of all measures when examining which measures predict accuracy for Experiment 1

Predictor	Model 15	Model 16	Model 17	Model 18	Model 19	Model 20	Model 21	Model 22	Model 23	Model 24
Fixted effects	Z-scored significant measures	Z-scored significant measures	Confidence age interaction	Box sorting age interaction	Response time age interaction	Head tilt age interaction	Hedges age interaction	Fillers age interaction	Booster age interaction	Final model
Intercept	1.52 (0.11)***	1.54 (0.09)***	1.60 (0.10)***	1.54 (0.09)***	1.54 (0.09)***	1.56 (0.09)***	1.52 (0.09)***	1.54 (0.09)***	1.55 (0.09)***	1.62 (0.10)***
Confidence	0.60 (0.08)***	0.62 (0.08)***	0.64 (0.08)***	0.60 (0.08)***	0.62 (0.08)***	0.62 (0.08)***	0.61 (0.08)***	0.62 (0.08)***	0.61 (0.08)***	0.64 (0.08)***
Box sorting	-0.05 (0.07)	-0.02 (0.07)	-0.02 (0.07)	-0.02 (0.07)	-0.09 (0.07)	-0.02 (0.07)	-0.01 (0.07)	-0.02 (0.07)	-0.02 (0.07)	0.01 (0.07)
Response time	-0.16 (0.07)*	-0.13 (0.07)	-0.11 (0.07)	-0.13 (0.07)	-0.13 (0.07)	-0.14 (0.07)*	-0.13 (0.07)	-0.13 (0.07)	-0.13 (0.07)*	-0.13 (0.07)
Head tilt	-0.13 (0.06)	-0.14 (0.06)	-0.14 (0.06)	-0.14 (0.06)*	-0.14 (0.06)*	-0.13 (0.06)*	-0.14 (0.06)*	-0.14 (0.06)	-0.14 (0.06)*	-0.14 (0.06)*
Hedges	-0.06 (0.09)	-0.06 (0.07)	-0.06 (0.07)	-0.06 (0.07)	-0.06 (0.07)	-0.06 (0.07)	-0.06 (0.08)	-0.06 (0.07)	-0.06 (0.07)	-0.06 (0.07)
Fillers	-0.03 (0.07)	-0.04 (0.07)	-0.03 (0.07)	-0.04 (0.07)	-0.04 (0.07)	-0.03 (0.07)	-0.03 (0.07)	-0.03 (0.07)	-0.04 (0.07)	-0.03 (0.07)
Boosters	0.15 (0.09)	0.15 (0.09)	0.13 (0.07)	0.14 (0.09)	0.15 (0.09)	0.14 (0.09)	0.15 (0.08)	0.15 (0.08)	0.18 (0.09)	0.13 (0.09)
Age		0.37 (0.09)***	0.45 (0.08)***	0.40 (0.09)***	0.37 (0.09)***	0.40 (0.09)***	0.37 (0.09)***	0.37 (0.09)***	0.39 (0.10)***	0.48 (0.10)***
Confidence x age			0.20 (0.07)**							0.17 (0.07)*
Box sorting x age				-0.11 (0.06)*						-0.06 (0.06)
Reaction time x age					0.001 (0.05)					
Head tilt x age						-0.12 (0.06)*				-0.11 (0.06) *
Hedges x age							0.01 (0.07)			
Fillers x age								-0.03 (0.07)		
Boosters x age									0.09 (0.11)	
Random parameters										
Level 2 intercept variance	0.26 (0.51)	0.10 (0.32)	0.08 (0.29)	0.10 (0.31)	0.10 (0.32)	0.11 (0.33)	0.10 (0.32)	0.10 (0.32)	0.09 (0.30)	0.09 (0.31)
Model fit										
Model df	9	10	11	11	11	11	11	11	11	13
Test change in df		1	1	1	1	1	1	1	1	2
AIC	1330.4	1323.3	1323.3	1328.4	1332.4	1327.8	1332.4	1332.3	1331.6	1322.0
BIC	1383.1	1381.3	1381.3	1386.4	1390.3	1385.7	1390.3	1390.2	1389.5	1390.5
-2 log likelihood	-655.2	-650.7	-650.7	-653.2	-655.2	-652.9	-655.2	-655.1	-654.8	-648

Asterisks indicate measures that significanly predcit accuracy, * p < 0.05, ** p < 0.01, *** p < 0.001.

2.5.5.5 Accuracy characteristic plots

Following the results from the regression, we plotted accuracy characteristic plots for the measures that significantly predicted accuracy. The graphs were noisy when separated by each year of age (e.g., aged 4-, 5-, 6- years), so we split the participants into 2 age groups. In line with the developmental literature that suggests that metacognitive development is believed to begin at the age of 4 years, and be robust around the age of 8 years, we split the participants into a younger (4-6 years) age group, and an older (7-8 years) age group. We plotted proportion correct for both age groups as a function of confidence, and as a function of each significant implicit measures (e.g., open or closed-eye for box sorting; presence or absence for body and vocal measures). Proportion correct was calculated by first calculating the hit rate (HR) and false alarm rate (FAR) by dividing both the number of hits (i.e., correct answers) and false alarms (i.e., number of incorrect answers) by the number of trials. Next, we divided the hit rate by the false alarm rate. In all plots, the dashed line indicates chance accuracy for uncertainty indicators (e.g., at the lowest confidence, closed-eye box, slowest response times, presence of head tilt, fillers, hedges, and absence of boosters) and perfect accuracy for certainty indicators (e.g., at the highest confidence, open-eye box, fastest response times, absence of head tilt, fillers, hedges, and presence of boosters). Error bars are standard errors.

2.5.5.5.1 Confidence-accuracy characteristic

As the data on the 5-point water cup scale were noisy, we collapsed the confidence ratings to make a 3-point scale (empty to ¼ full, ½ full to ¾ full, completely full) to better visualise the data (see also Winsor et al., in press). Figure 3A indicates that confidence was predictive of accuracy for both age groups, but more predictive for the older children. Older children’s high confidence responses were 97% correct, and their low confidence responses were 62% accurate. The younger children were still able to assign higher confidence judgements to their correct answers, and lower judgements to their incorrect answers; their high confidence responses were 83% accurate, and their low confidence responses 56% correct. This explains the significant interaction between confidence and age in the modelling and indicates that the informativeness of confidence for predicting memory accuracy improves with age.

2.5.5.5.2 Box sorting-accuracy characteristic

Figure 3B indicates that box sorting was predictive of accuracy in older children, but not for younger children. Older children were adept at sorting their accurate answers into the open-eye box and their inaccurate responses into the closed-eye box; responses sorted into the open-eye box were 86% correct, whist responses in the closed-eye box were only 58% correct. Younger children, however, achieved similar accuracy for answers in the open-eye box (77%) and the closed-eye box (74%). This explains the significant interaction between box sorting and age in the modelling and indicates that the informativeness of box sorting in predicting memory accuracy improves with age.

2.5.5.5.3 Response time-accuracy characteristic

We created three response time groups: mid response times were coded as between 5 and 10 seconds (*M* = 6.83, *SD* = 1.62), with fast and slow response times falling either side of these times (Fast *M* = 2.41, *SD* = 1.09 Range = 4; Slow *M* = 16.59, *SD* = 12.44, Range = 93). Figure 3C indicates that response time was predictive of accuracy in both older and younger children. For older children, fast responses given within 4 seconds were 90% correct, mid responses were 76% correct, and slow responses given after 10 seconds were 68% correct. For younger children, fast responses made within 4 seconds were 83% correct mid responses were 75%, and slow responses made after 10 seconds were 63% correct. Older children achieved a higher proportion of correct answers within the fast time frame than younger children.

2.5.5.5.4 Head tilt-accuracy characteristic

Figure 3D indicates that head tiles were predictive of accuracy in older children, but less so for younger children. For older children, answers that were not accompanied with a head tilt were 85% correct, whereas answers that were accompanied with a head tilt were only 61% correct. For younger children the results were in the predicted direction, but the effect was small and did not appear to be reliable because the error bars overlap: answers that were not accompanied with a head tilt were 77% correct, whereas answers that were accompanied with a head tilt were 70% correct. This explains the significant interaction between head tilt and age in the modelling and indicates that the informativeness of head tilts for predicting memory accuracy improves with age.

2.5.5.5.5 Hedges-accuracy characteristic

Figure 3E indicates that hedges were predictive of accuracy for both older and younger children, but slightly stronger in younger children. For older children, answers were 84% correct when hedges were absent, and 75% correct when hedges were present. The relationship was slightly stronger in the younger children, with a 15% difference in accuracy on average between answers when hedges were absent and present; answers were 78% correct when hedges were absent and 63% correct hedges were present. Older children achieved a higher proportion correct when hedges were both absent and present compared to younger children, but the plot and the modelling suggest that hedges were predictive of accuracy for both younger and older children.

2.5.5.5.6 Filler-accuracy characteristic

Figure 3F indicates that fillers were predictive of accuracy for both older and younger children. For older children, answers were 84% correct when fillers were absent and 73% correct when fillers were present. Similarly, for young children answers were 78% correct when fillers were absent and 65% correct when fillers were present. Older children achieved a higher proportion of correct answers when fillers were absent than younger children, but the plot and the modelling suggest that fillers were predictive of accuracy for both younger and older children.

2.5.5.5.7 Booster-accuracy characteristic

Figure 3G indicates that boosters were predictive of accuracy for both older and younger children. For older children, answers were 95% correct when boosters were present, and 82% correct when boosters were absent. For younger children answers were 92% correct when boosters were present, and 84% correct when boosters were absent. Therefore, the plot and the modelling suggest that boosters were predictive of accuracy for both younger and older children.

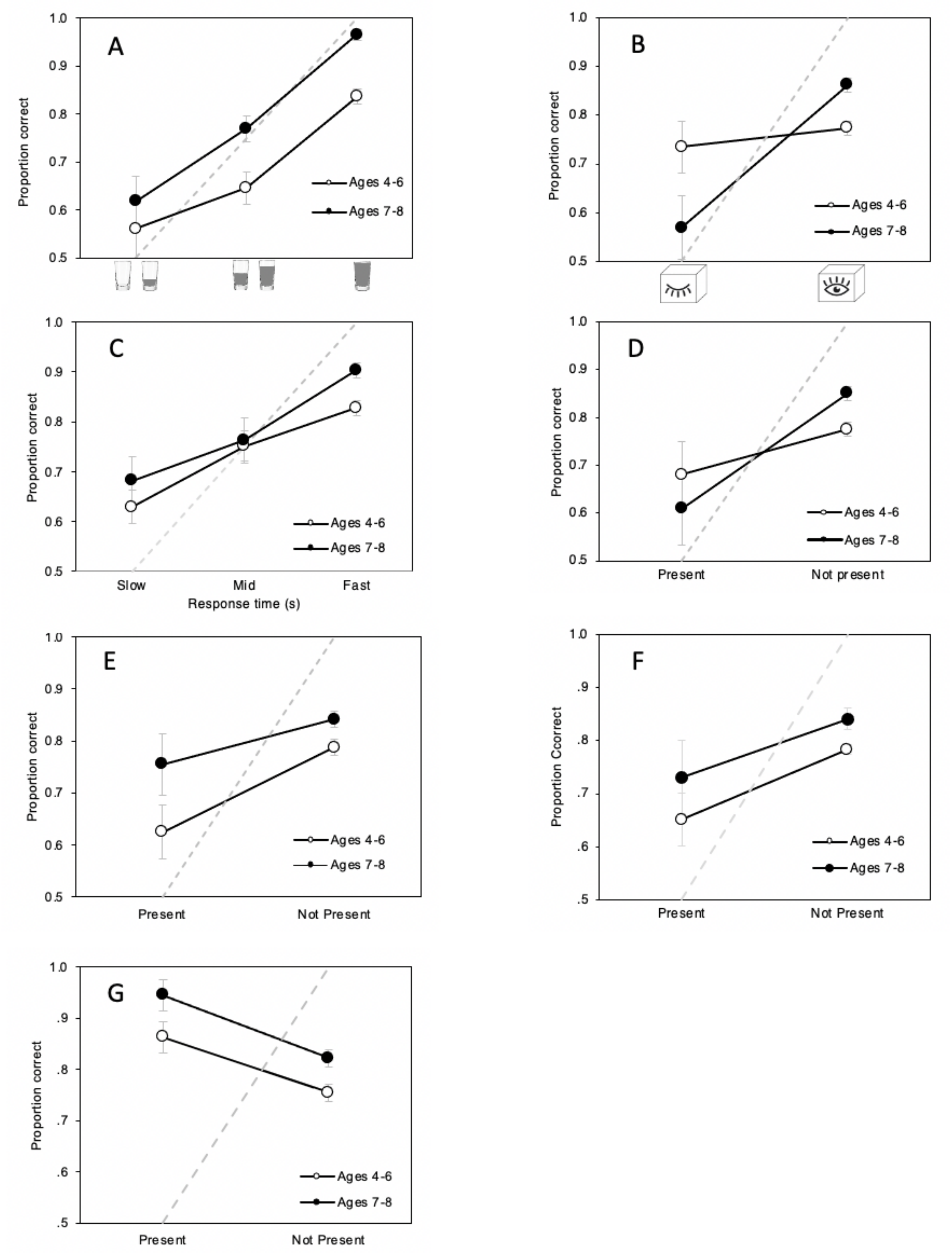


Figure 3. Accuracy characteristic plots for the 7 significant measures in Experiment 1 (A) Confidence, (B) Box sorting, (C) Response time, (D) Head tilts, (E) Fillers, (F) Hedges, (G) Boosters. The dashed line represents chance-level performance.

2.6 Summary

In sum, the aim of Experiment 1 was to investigate which implicit and explicit metacognitive measures were predictive of accuracy in children aged 4-8 years, and how the informativeness of these measures change with age. The results from Experiment 1 indicate that confidence, box sorting, response time, head tilts, hedges, fillers, and boosters were predictive of accuracy. Confidence, head tilts and response time were uniquely predictive of accuracy when controlling for other measures, with confidence being the most predictive measure. When age was added as a main effect, age also became a unique predictor, but response time was no longer a unique predictor, indicating that the unique variance explained by response time was accounted for by age. Finally, the informativeness of confidence, box sorting and head tilts changed with age: they all became more informative with age, with confidence being the measure that improved most with age. Overall, these results suggests that there are both explicit (e.g., confidence) and implicit measures (e.g., response time, box sorting, head tilts, hedges, fillers, and boosters) that could be used predict memory accuracy in children aged 4-8 years, but that confidence is the most informative measure. A pre-registered replication of Experiment 1 was conducted (i.e., Experiment 2) to examine if the findings from Experiment 1 replicated in another sample of children aged 4-8 years.

Moreover, one of the key findings from Experiment 1 was that both younger and older children had a confidence-accuracy relationship, despite confidence being an explicit measure. Perhaps most importantly, confidence was more informative of memory accuracy than age. This result suggests that confidence ratings from children reflect their likely accuracy, and that these ratings should be considered over a child's age when considering their memory evidence. This could be relevant in a legal setting, as currently a 5-year old's eyewitness account may be disregarded, but an 8-year-old's account not disregarded, based on the belief that the 5-year-old has poorer memory, and cannot identify when their memories are accurate (e.g., Keast, Brewer, & Wells, 2007; Powell, Garry, & Brewer, 2013). Our results suggest that, although 5-year old's may have poorer memory than older children and adults, they are able to indicate their level of confidence that accurately corresponds to their memory strength. This result is reflected in previous research (e.g., Hembacher & Ghetti, 2014) and a reanalysis (Winsor et al., in press).

Interestingly, the results of the box-sorting task (an implicit measure) indicated that younger children had no box sorting-accuracy relationship, whilst older children did. This indicates that, whilst younger children were able to monitor their uncertainty and express this through confidence judgements, there were not able to strategically control their behaviour when deciding which answers to share with the researcher. As discussed in the introduction, this could be due to the linguistic complexity of the task, with younger children having more difficulty understanding than older children (Darnell, 2015; Smith, Shields, & Washburn, 2003; Pratt & Bryant, 1990). Beck, Robinson, & Freeth (2008) found that children of a similar age (5-6 years) struggled to implement a delay strategy in place of answering a question they were uncertain of, perhaps indicating difficulty in understanding the task. Older children had a relationship between box sorting and accuracy, indicating they were able to control their behaviour, in accordance with their monitoring. This is in line with literature that suggests a developmental increase in control ability (e.g., Istomina, 1975;1982).

Nevertheless, one limitation of the current experiment was that children always provided a confidence rating before their box sorting decision (following Hembacher & Ghetti, 2014). Therefore Experiment 2 will further explore the relationship between confidence and box sorting and examine whether having confidence as a precursor to box sorting facilitated children's box sorting decisions. Flavell, Green, & Flavell (1995) stated that active reflection on mental states can aid in subsequent behavioural strategy. It could be that by asking the children how sure they were of their answer encouraged active reflection on their uncertainty, and in turn helped in making a box sorting decision (Brown & Walker, 1983; Fisher, 1998). Younger children may have performed more poorly on the box sorting task as they are less able to actively reflect upon their uncertainty. If confidence does facilitate box sorting decision, then, based on the results from Experiment 1, we would expect a weaker relationship between box sorting and accuracy in older children in the confidence absent condition. We would also expect younger children to continue to have no box sorting-accuracy relationship, as the results for Experiment 1 suggests that their control behaviour does not benefit from active reflection.

CHAPTER 3

EXPERIMENT 2

3.1 Introduction

The aim of Experiment 2 was to attempt to replicate the findings from Experiment 1 in another sample of children aged 4-8 years. This was an important step in avoiding errors of inference from Experiment 1, as if a well powered second study yielded the same results from the first study, more stable conclusions can be drawn from the results (Asendorpf & Baudonnière, 1993). A secondary aim was to further explore the relationship between monitoring and control in children aged 4-8 years by examining if box-sorting decisions are more informative of accuracy when preceded with a confidence judgement. Therefore, an additional research question concerning whether children are still able to efficiently implement control processes (i.e., box sorting) without an explicit monitoring component (i.e., confidence judgements) was considered in Experiment 2.

In Experiment 1, older children (7-8 years) were able to both appropriately rate their confidence to reflect their accuracy and sort their answers into the boxes according to accuracy. Together, this performance suggests that they were able to monitor and reflect on their uncertainty and implement strategy accordingly. Younger children (4-6 years) were able to do the former, but not the latter, possibly suggesting that they can monitor their memory, but not control their behaviour (Figure 3B). The aim of Experiment 2 was to explore if confidence judgements facilitated the appropriate control behaviours in older children, and if older children would still be able to use the boxes in accordance with certainty without explicitly stating their monitoring first through a confidence judgement.

Another point of interest was why younger children did not perform well on the box sorting task (an implicit measure) but performed well on the confidence task (an explicit measure) in Experiment 1. This result was surprising, in that whilst research suggests younger children can indicate their uncertainty, that this is done implicitly. It has been demonstrated that younger children struggle with explicit indicators of certainty (i.e., confidence judgements). Additionally, younger children are reported to have trouble using scales (Chambers & Johnston, 2002).

As it stands in Experiment 1, younger children seemed to be able to monitor their uncertainty and report this explicitly but were unable to control their behaviour to reflect their accuracy, suggesting a disconnect between monitoring and control. It could be that whilst they appear able to reflect upon their uncertainty and express this explicitly through confidence judgements, they are not yet able to use this information to inform their control behaviour through a subsequent strategy. This would be in line with both the *availability deficiency* (Veenman, Kerseboom , & Imthorn, 2000; Winne, 1996) hypothesis, where children do not have skills beyond monitoring to control their behaviour, *and the production deficiency* (Flavell, Beach & Chinksy, 1966) hypothesis, where children have the control skills, but do not understand how the box sorting task will utilise this skill.

Given theories of metacognitive monitoring and control described in section 1.2.5, and previous research on these theories (e.g., that monitoring aspects are integral to control implementation), we were interested to see if the box sorting task was better utilised by the children when confidence judgements are a precursor. If children's control is facilitated by explicit monitoring, then we would expect children to perform better on the box sorting task when confidence is present. In other words, assigning answers to the appropriate box will be facilitated by the confidence judgement given before as it requires children to reflect upon and report their level of certainty (e.g., Gill, Swann, & Silvera, 1998; Nelson & Narens, 1990; Son & Schwartz, 2002). If younger children are unable to use the box sorting task in this second study, then this will add to evidence of children under the age of 7 years being less able to use monitoring abilities to guide their behaviour than older children (Flavell, Beach & Chinksy,1966).

To achieve these aims, Experiment 1 was repeated, but with the added between-subjects conditions: confidence present (confidence rating scale preceded the box-sorting task) or confidence absent (no confidence rating scale preceded the box-sorting task). It was also noted in Experiment 1 that some children stopped pointing during the Experiment and verbalised their answers instead. To ensure that children continued to point at their answers throughout the duration of the study, the instructions in Experiment 2 were adjusted slightly so that each question began with "can you point to…".

3.2 Experiment 2: Method

3.2.1 Design

We used a 2 (age: younger, older children) x 2 (confidence condition: confidence present, confidence absent) between-subject design. Video order and confidence condition were counterbalanced. Question order was randomised for each participant. Our data-collection stopping rule was to recruit at least 80 children. The research was reviewed according to the University of Birmingham Science, Technology, Engineering and Mathematics Ethical Review Committee.

3.2.2 Participants

A total of 106 children between the ages of 4-8 years were recruited via advertising on social media platforms (e.g., Facebook, Twitter) and recruitment websites (childrenhelpingscience.com, callforparticipants.com, honeybee.io, psych.hanover.edu). This sample size was selected to ensure statistically adequate power for a between-subject experiment, and to further avoid type 1 (false positive, α) and type 2 (false negative, β) errors (Asendorpf, et al., 2013; Maxwell, Kelley, & Rausch, 2008). From the 106 children collected, data from 20 children were excluded due to adult interference and technical difficulties (e.g., poor internet connection, poor video quality, issues with Zoom). This resulted in a final sample of 86 participants. (*M* age = 6.22; *SD* age = 1.27; 57% female). Out of the 86 included participants, an additional 197 trials were excluded from 26 children due to adult interference and deviation from the instructions. Therefore, a total of 917 trials were removed from the dataset of 86 participants. A further four trials were removed due to missing box sorting data, leaving 3298 completed trials to be analysed.

3.2.3 Materials

Materials were identical to Experiment 1 (see section 2.5.3).

3.2.4 Procedure

The task was identical to Experiment 1, except for this time children were assigned to either a confidence present or confidence absent condition. Children in the confidence present condition were given the same instructions as Experiment 1 and reported their confidence judgement and box sorting decision (see Appendix A for experiment set up). Children in the confidence absent condition completed the box sorting task, but not the confidence rating task (see Appendix B for experiment set up). Children were asked to point to their answers (e.g., "can you point to which hat the boy was wearing?").

3.2.5 Results

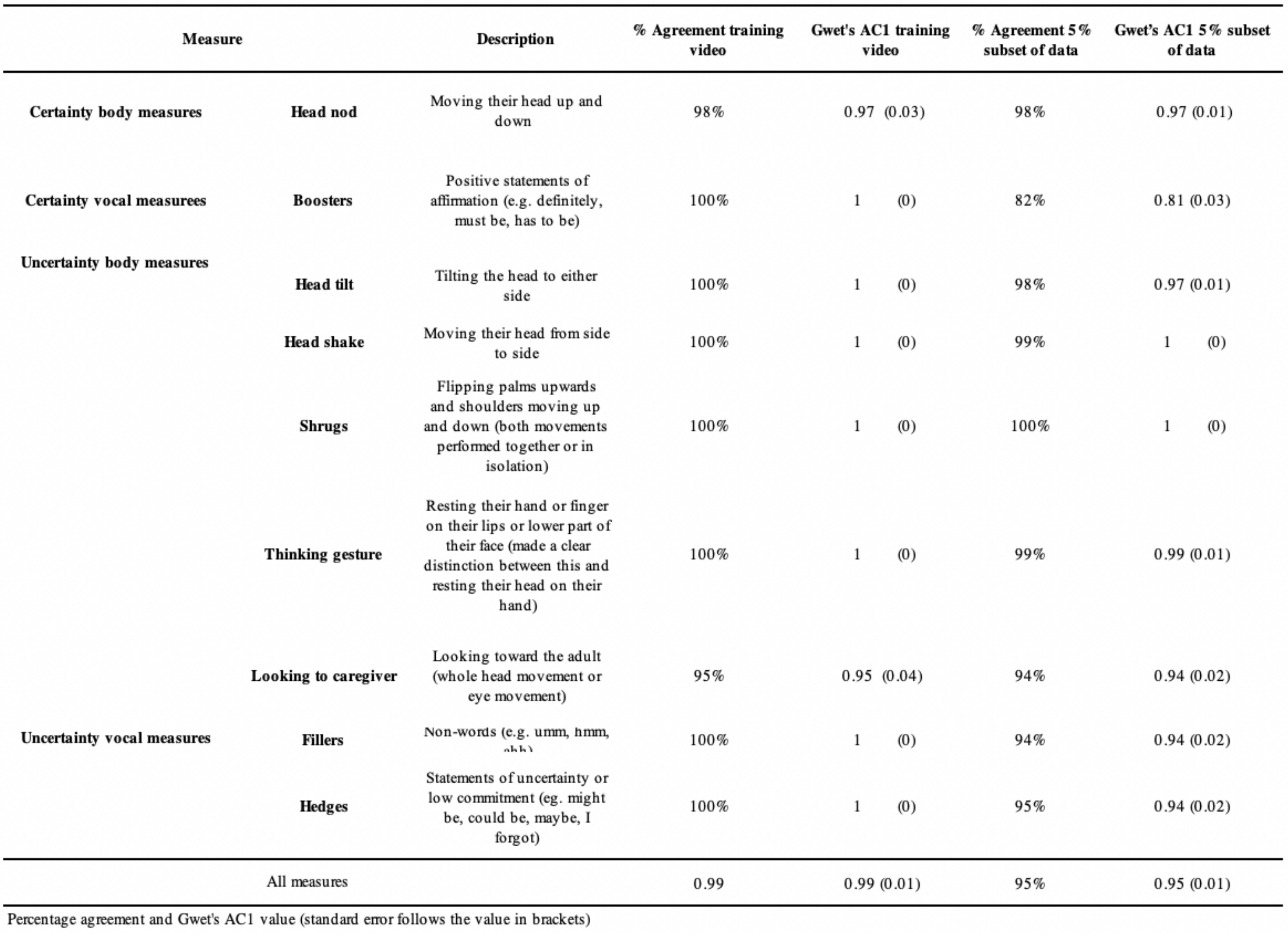
3.2.5.1 Memory Accuracy

There were 44 younger children (aged 4-6, *M* age = 5.16, *SD* = .81) and 42 older children (aged 7-8, *M* age = 7.33, *SD* = .48). Performance on the memory task in Experiment 2 was high for all ages (*M* = .81, *SD* = 40). Younger children's mean accuracy (*M* = .75, *SD* = .43) was -.12 (*SE* = .02) than older children's mean accuracy (*M* = .86, *SD* = .34). An independent-samples t-test was run to determine if there were significant differences in accuracy between younger and older children. The difference in mean accuracy between age groups was statistically significance, *t*(62.37) = 6.56, *p* = <.001.

3.2.5.2 Coding

The coding protocol from Experiment 1 was used. Two coders blind to the purpose of the study coded the implicit measures displayed in the videos by the child participants. The two coders also coded the implicit measures in Experiment 1. Each coder individually coded the testing session of one participant who completed 40 trials (i.e., completed a video for training purposes). There was a high percentage of agreement between coders (99% overall). For the training video coded, the overall interrater reliability across the two coders was AC_1= .99, indicating very good reliability (Gwet, 2014; see section 2.5.5.2, Chapter 2). The Gwets AC_1 results for each implicit measure coded in the training video is shown in Table 6.

As the reliability of coding was sufficiently high, once training was complete, each coder independently coded the data for approximately 50 participants. In addition, both coders coded the same 6% of participants (N = 5 participants) and we again measured the interrater reliability, which was very good overall (AC_1 = .95). The interrater reliability for all individual measures were also calculated and are presented in Table 6. In short, these results illustrate that the two coders agreed about the measures that they coded.



Measure		Description	% Agreement training video	Gwet's AC1 training video	% Agreement 5% subset of data	Gwet's AC1 5% subset of data
Certainty body measures	**Head nod**	Moving their head up and down	98%	0.97 (0.03)	98%	0.97 (0.01)
Certainty vocal measurees	**Boosters**	Positive statements of affimation (e.g. definitely, must be, has to be)	100%	1 (0)	82%	0.81 (0.03)
Uncertainty body measures	**Head tilt**	Tilting the head to either side	100%	1 (0)	98%	0.97 (0.01)
	Head shake	Moving their head from side to side	100%	1 (0)	99%	1 (0)
	Shrugs	Flipping palms upwards and shoulders moving up and down (both movements perfomed together or in isolation)	100%	1 (0)	100%	1 (0)
	Thinking gesture	Resting their hand or finger on their lips or lower part of their face (made a clear distinction between this and resting their head on their hand)	100%	1 (0)	99%	0.99 (0.01)
	Looking to caregiver	Looking toward the adult (whole head movement or eye movement)	95%	0.95 (0.04)	94%	0.94 (0.02)
Uncertainty vocal measures	**Fillers**	Non-words (e.g. umm, hmm, ahh)	100%	1 (0)	94%	0.94 (0.02)
	Hedges	Statements of uncertainty or low commitment (eg. might be, could be, maybe, I forgot)	100%	1 (0)	95%	0.94 (0.02)
	All measures		0.99	0.99 (0.01)	95%	0.95 (0.01)

Percentage agreement and Gwet's AC1 value (standard error follows the value in brackets)

Table 6. The percentage of agreement and Gwet's AC1 reliability score of all coded measures for training video and 5% subset of the data for Experiment 2

3.2.5.3 Z-transformed measures

To visualise our data, we Z-transformed the mean amount of each measure and plotted them as a function of accuracy (Gustafsson, Lindholm, & Jönsson, 2019). Descriptively speaking, Figure 4 shows that when children were incorrect, they made lower confidence judgments, responded more slowly, and more often chose to hide their answers on the box sorting task. Considering implicit measures, they performed more thinking gestures, head shakes, looking to caregiver, hedges, and fillers. They performed few head tilts, and shrugs. When children were correct, they made higher confidence judgements, responded more quickly, and more often chose to show their answers on the box sorting task. considering implicit measures, they performed more boosters, but performed fewer head nods.

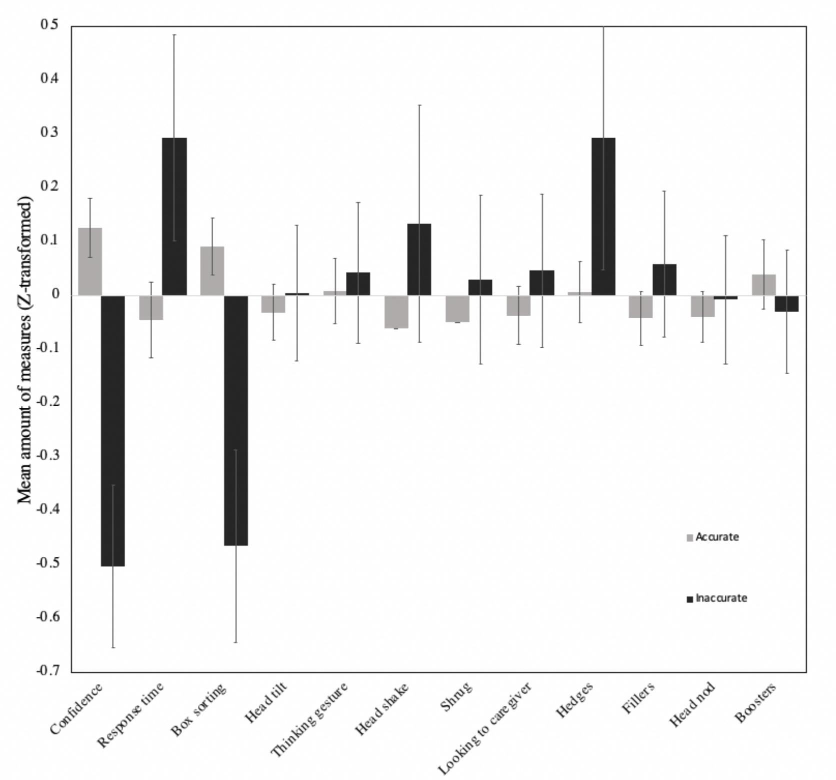


Figure 4. Mean amount of implicit and explicit measures (z-transformed) for accurate and inaccurate answers for Experiment 2. Error bars represent 95% confidence intervals

3.2.5.4 Multilevel modelling

Next, to answer how predictive the measures were of accuracy, we fitted a series of multilevel logistic regression models using the lme4 package in R. As in Experiment 1, we organised the data as multilevel data with individual responses nested within participants and followed the methods used in Gustafsson, Lindholm, & Jönsson (2019) and Mansour, Beaudry, & Lindsay (2017). To assess model fit, we again used the likelihood ratio test.

Participants in the confidence-absent condition did not provide confidence judgements. As a result, 59.3% of the final dataset contained planned missing data (i.e., after data exclusions there were 51 children in the confidence-absent condition), and so a multiple imputation was conducted in R using a Random Forests based method in the MICE package (van Buuren & Groothuis-Oudshoorn, 2011; Rubin, 1987). The mean confidence score for the imputed data (M = 4.19, SD = 1.17), was identical to the mean confidence score for the observed data (M = 4.19, SD = 1.18), indicating the imputation was successful in producing realistic results. (Nguyen, Carlin, & Lee, 2017).

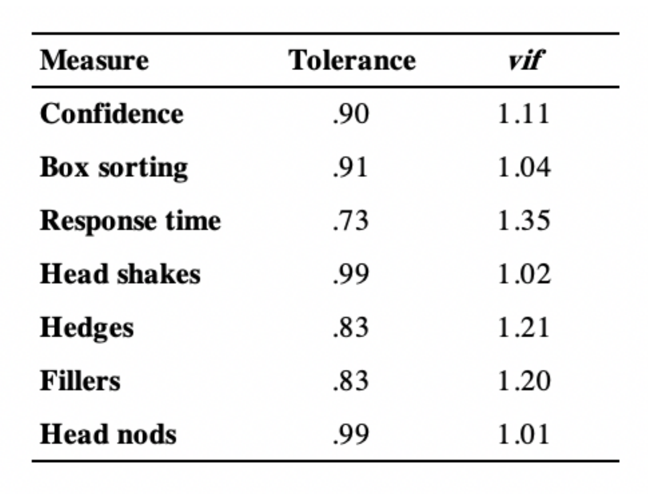
3.2.5.4.1 Examining which measures predict memory accuracy

First, to examine which of the measures were significantly predictive of accuracy, we compared models for each of the 12 measures (models 2-13) to a null intercept-only model predicting accuracy (model 1). Table 9 reports the model parameters and fit indices for models 1-14. Note that tests for skewness indicated that some of the data were not normally distributed.

Compared to the null model (model 1, intercept), the model fit was significantly improved when adding confidence, $\chi^2(1) = 171.63$, $p > .001$ (model 2); box sorting, $\chi^2(1) = 62.10$, $p > .001$ (model 3); response time, $\chi^2(1) = 74.93$, $p >.001$ (model 4); head shakes,

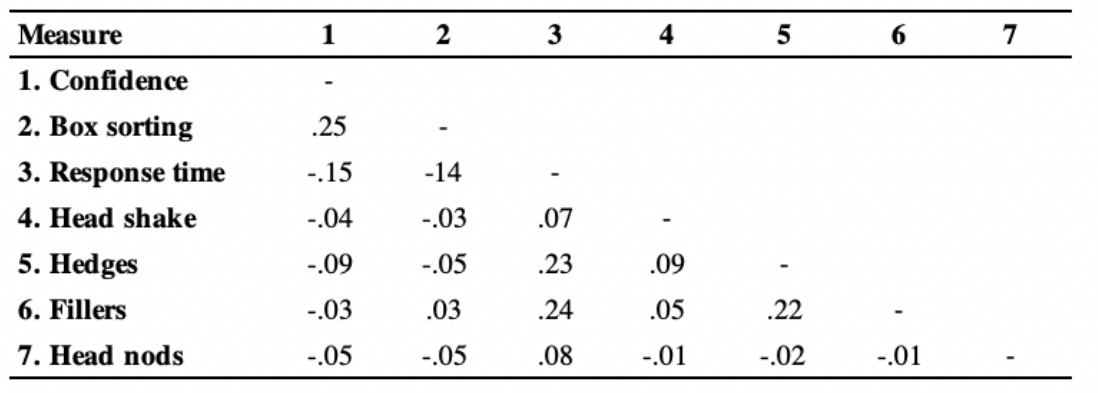
$\chi^2(1) = 5.87$, $p = .02$ (model 6); hedges, $\chi^2(1) = 10.77$, $p = .001$ (model 10); fillers, $\chi^2(1) = 4.37$, $p = 0.04$ (model 11); and head nods, $\chi^2(1) = 5.81$, $p = .02$ (model 12); but not by head tilts, $\chi^2(1) = .31$, $p = .0.58$ (model 5); shrugs, $\chi^2(1) = 1.37$, $p = .24$ (model 7); thinking gestures, $\chi^2(1) = 1.45$, $p = .23$ (model 8); looking to caregiver, $\chi2(1) = 3.08$, $p = .08$ (model 9); or boosters, $\chi^2(1) = 0.29$, $p = .59$ (model 13). As with Experiment 1, tests to see if the data for the 7 significant predictors met the assumptions of collinearity indicated that multicollinearity was not a concern. Tolerance and *vif* values for the 7 significant predictors indicated no incidence of multicollinearity (see Table 7 & Table 8).

Table 7. Significant measures tolerance and *vif* statistic for Experiment 2



Measure	**Tolerance**	***vif***
Confidence	.90	1.11
Box sorting	.91	1.04
Response time	.73	1.35
Head shakes	.99	1.02
Hedges	.83	1.21
Fillers	.83	1.20
Head nods	.99	1.01

Table 8. Significant measures correlation matrix for Experiment 2



Measure	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. Confidence	-						
2. Box sorting	.25	-					
3. Response time	-.15	-14	-				
4. Head shake	-.04	-.03	.07	-			
5. Hedges	-.09	-.05	.23	.09	-		
6. Fillers	-.03	.03	.24	.05	.22	-	
7. Head nods	-.05	-.05	.08	-.01	-.02	-.01	-

The 7 significant predictors were added to one model (model 14), which was then compared to each significant predictor model separately. The model with all 7 significant predictors (model 14) improved the fit relative to the confidence model, $\chi^2(6) = 60.59$, p = >.001 (model 2); box sorting model, $\chi^2(6) = 170.12$, $p < .001$ (model 3); response time model, $\chi^2(6) = 157.30$, $p < .001$ (model 4); head shakes model, $\chi^2(6) = 226.36$, $p < .001$ (model 6); hedges model, $\chi^2(6) = 221.45$, $p < .001$ (model 10); fillers model, $\chi^2(6) = 227.86.42$, $p < .001$ (model 11); and head nods model, $\chi^2(6) = 226.41$, $p < .001$ (model 12), indicating that a model with all significant predictors better predicted accuracy than each of the significant predictors alone.

3.2.5.4.2 Examining which measures uniquely predict memory accuracy

Next, we z-transformed each of the significant measures and used the z-transformed values to examine which measures uniquely predicted accuracy when controlling for the other predictors. We fit the model with all 7 significant predictors (model 14) using the z-transformed measures (model 15). Table 10 reports the model parameters and fit indices for models 15-24 using the z-transformed measures. Replicating Experiment 1, confidence (z = 6.83, $p < .001$) and response time ($z = -5.59$, $p < .001$) were uniquely predictive of accuracy when controlling for the other predictors. In Experiment 2, box sorting ($z = 3.62$, $p = .003$) was also a unique predictor of accuracy. Specifically, as accuracy increased, confidence increased, response time decreased, and children were more likely to show their answers. As in Experiment 1, fillers ($z = .67$, $p = .50$); head shakes ($z = -1.11$, $p = .27$); hedges ($z = .73$, p = .46); and head nods ($z = -1.26$, $p < .21$) did not uniquely explain memory accuracy when controlling for the other predictors. This suggests that, although a model with all predictors predicts accuracy better than any of the significant predictors alone, the unique predictors each contribute a percentage of the variance that can be captured only by them alone. That is, the collection of confidence, response time, and box-sorting provides unique information to the prediction of accuracy over and above the other significant measures. Conversely, the collection of fillers, head shakes, hedges and head nods does not provide unique information can cannot be captured by the other measures.

3.2.5.4.3 Examining which measures predict memory accuracy better than age

We next examined if any measures were more predictive of accuracy than age. We took the z-transformed model with 7 predictors (model 15) and added age as a main effect (model 16). Age was also z-transformed. Adding age as a main effect significantly improved the model fit, indicating that memory accuracy improved with age. We now refer to this model as the best-fitting main effects model. Confidence (z = 10.02, p < .001), box sorting (z = 3.31, p < .001), response time (z = -4.85, p < .001) and age (z = 6.47, p < .001) all uniquely explained memory accuracy when controlling for the other predictors. As accuracy increased, confidence increased, open-eye box decisions increased, response time decreased, and age increased. Additionally, as in Experiment 1, the size of the z scores indicates that confidence was more informative of accuracy than box sorting, response time and age.

3.2.5.4.4 Examining if the informativeness of the measures change with age

To examine if the informativeness of the measures changed with age, we investigated if adding the fixed interaction of age separately with each significant predictor (model 17-23) would improve the model fit compared to the best-fitting main effects model (model 16). In Experiment 1, the interaction between age and confidence, age and box sorting, and age and head tilts significantly improved the fit of the model. As in Experiment 1, adding the interaction between age and confidence, χ^2 (11) = 12.16, *p* = <.001 (model 17); and age and box sorting, χ^2(11) = 5.36, *p* = .02 (model 18) improved the model fit. Additionally, this time, the interaction between age and response time, χ^2(11) = 48.39, *p* = <.001 (model 19); age and hedges, χ^2 (11) = 6.48, *p* = .01 (model 21); and age and fillers, χ^2 (1) = 6.68, *p* = .01 (model 22); also improved the model fit. This suggests that the informativeness of confidence, box sorting, response time, fillers, and hedges and in predicting memory accuracy changes with age. Adding the interaction between age and head shake, χ2(11) = 2.20, *p* = .14 (model 20); and age and head nods, χ^2(11) = 2.38, *p* = .12 (model 23) did not improve the model fit compared to the best-fitting main effects model. This suggest that the informativeness of head shakes and head nods in predicting memory accuracy does not change significantly with age.

Table 9. Model parameters of all measures when examining which measures predict accuracy

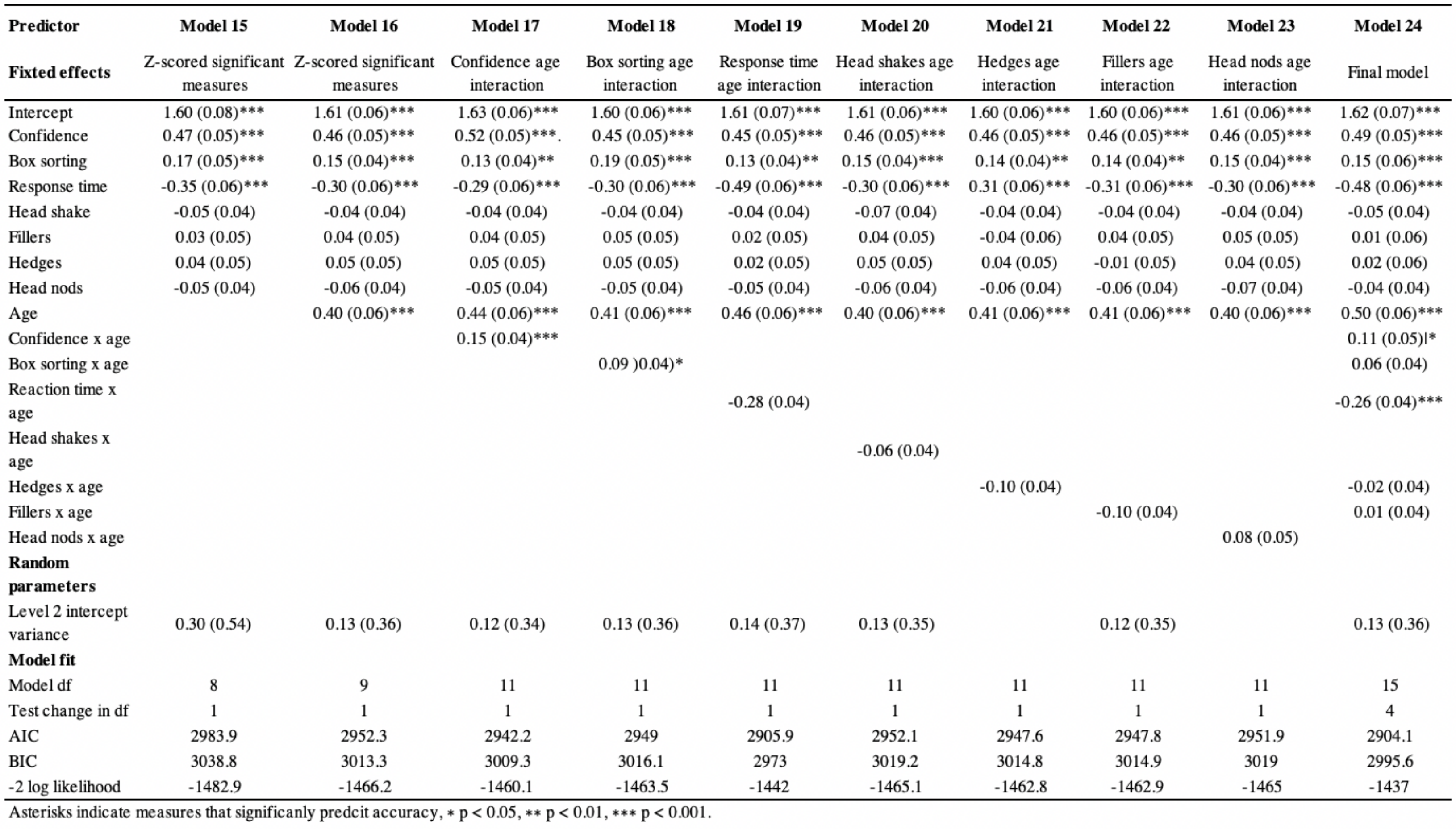
Predictor	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14
Fixted effects	Intercept	Confidence	Box sorting	Response time	Head tilt	Head shake	Shrugs	Thinking gesture	Looking to caregiver	Hedges	Fillers	Head nods	Boosters	Total model
Intercept	1.48 (0.07)***	-0.44 (0.16)**	0.44 (0.14)**	2.00 0.10***	1.48 (0.07)***	1.49 (0.07)***	1.48 (0.07)***	1.49 (0.07)***	1.50 (0.07)***	1.50 (0.07)***	1.51 (0.07)***	1.49 (0.07)***	1.49 (0.07)	-0.21 (0.21)
Confidence		0.48 (0.04)***												0.40 (0.04)***
Box sorting			1.17 (0.14)***											0.58 (0.16)***
Response time				-0.09 (0.01)***										-0.08 (0.01)***
Head tilt					-0.24 (0.42)									
Head shake						-1.52 (0.60)*								-0.76 (0.69)
Shrugs							-0.92 (0.75)							
Thinking gesture								-0.30 (0.24)						
Looking to caregiver									-0.38 (0.21)					
Fillers										-0.30 (0.14)*				0.12 (0.17)
Hedges											-0.50 (0.15)**			0.14 (0.20)
Head nod												-1.06 (0.42)*		-0.57 (0.46)
Boosters													-0.08 (0.15)	
Random parameters														
Level 2 intercept variance	0.24 (0.48)	0.30 (0.55)	0.19 (0.44)	0.25 (0.50)	0.24 (0.49)	0.24 (0.49)	0.23 (0.48)	0.24 (0.49)	0.24 (0.49)	0.23 (0.48)	0.23 (0.48)	0.24 (0.49)	0.24 (0.49)	0.30 (0.54)
Model fit														
Model df	2	3	3	3	3	3	3	3	3	3	3	3	3	8
Test change in df		1	1	1	1	1	1	1	1	1	1	1	1	5
AIC	3202.1	3032.5	3142.00	3129.20	3203.80	3198.30	3202.80	3202.70	3201.00	3199.80	3193.30	3198.30	3203.80	2983.90
BIC	3214.3	3050.8	3160.30	3147.50	3222.10	3216.60	3221.10	3221.00	3219.30	3218.10	3211.60	3216.60	3222.10	3038.80
-2 log likelihood	-1599.1	-1513.2	-1568.0	-1561.6	-1598.9	-1596.1	-1598.4	-1598.3	-1597.5	-1596.9	-1593.7	-1596.2	-1598.9	-1482.9

Asterisks indicate measures that significanly predcit accuracy, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Finally, to examine if a model with multiple interactions with age (i.e., confidence, response time, box sorting, hedges, and fillers) was better than models with one interaction, we took the best-fitting main effects model (model 16) and added interactions between age and confidence, age and box sorting, age and response time, age and hedges, and age and fillers. The model with all 5 interactions (model 24) was better than the model with the interaction between age the interaction between age and confidence (model 17), age and box sorting (model 18), age and hedges (model 21), and age and fillers (model22). The model with all 5 interactions (model 24) was a better fit compared to each individual age interaction model. Therefore, the final best fitting interaction model is model 24, which includes confidence, head nod, head shake, box sorting, fillers, hedges, response time and age as main effects, and the interaction between age and confidence, age and box sorting, age and response time, age and hedges, and age and fillers.

As with Experiment 1, we further visualise the interaction findings using accuracy characteristic plots.

Table 10. Model parameters of all measures when examining which measures predict accuracy



Predictor	Model 15	Model 16	Model 17	Model 18	Model 19	Model 20	Model 21	Model 22	Model 23	Model 24
Fixted effects	Z-scored significant measures	Z-scored significant measures	Confidence age interaction	Box sorting age interaction	Response time age interaction	Head shakes age interaction	Hedges age interaction	Fillers age interaction	Head nods age interaction	Final model
Intercept	1.60 (0.08)***	1.61 (0.06)***	1.63 (0.06)***	1.60 (0.06)***	1.61 (0.07)***	1.61 (0.06)***	1.60 (0.06)***	1.60 (0.06)***	1.61 (0.06)***	1.62 (0.07)***
Confidence	0.47 (0.05)***	0.46 (0.05)***	0.52 (0.05)***.	0.45 (0.05)***	0.45 (0.05)***	0.46 (0.05)***	0.46 (0.05)***	0.46 (0.05)***	0.46 (0.05)***	0.49 (0.05)***
Box sorting	0.17 (0.05)***	0.15 (0.04)***	0.13 (0.04)**	0.19 (0.05)***	0.13 (0.04)**	0.15 (0.04)***	0.14 (0.04)**	0.14 (0.04)**	0.15 (0.04)***	0.15 (0.06)***
Response time	-0.35 (0.06)***	-0.30 (0.06)***	-0.29 (0.06)***	-0.30 (0.06)***	-0.49 (0.06)***	-0.30 (0.06)***	0.31 (0.06)***	-0.31 (0.06)***	-0.30 (0.06)***	-0.48 (0.06)***
Head shake	-0.05 (0.04)	-0.04 (0.04)	-0.04 (0.04)	-0.04 (0.04)	-0.04 (0.04)	-0.07 (0.04)	-0.04 (0.04)	-0.04 (0.04)	-0.04 (0.04)	-0.05 (0.04)
Fillers	0.03 (0.05)	0.04 (0.05)	0.04 (0.05)	0.05 (0.05)	0.02 (0.05)	0.04 (0.05)	-0.04 (0.06)	0.04 (0.05)	0.05 (0.05)	0.01 (0.06)
Hedges	0.04 (0.05)	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)	0.02 (0.05)	0.05 (0.05)	0.04 (0.05)	-0.01 (0.05)	0.04 (0.05)	0.02 (0.06)
Head nods	-0.05 (0.04)	-0.06 (0.04)	-0.05 (0.04)	-0.05 (0.04)	-0.05 (0.04)	-0.06 (0.04)	-0.06 (0.04)	-0.06 (0.04)	-0.07 (0.04)	-0.04 (0.04)
Age		0.40 (0.06)***	0.44 (0.06)***	0.41 (0.06)***	0.46 (0.06)***	0.40 (0.06)***	0.41 (0.06)***	0.41 (0.06)***	0.40 (0.06)***	0.50 (0.06)***
Confidence x age			0.15 (0.04)***							0.11 (0.05)|*
Box sorting x age				0.09)0.04)*						0.06 (0.04)
Reaction time x age					-0.28 (0.04)					-0.26 (0.04)***
Head shakes x age						-0.06 (0.04)				
Hedges x age							-0.10 (0.04)			-0.02 (0.04)
Fillers x age								-0.10 (0.04)		0.01 (0.04)
Head nods x age									0.08 (0.05)	
Random parameters										
Level 2 intercept variance	0.30 (0.54)	0.13 (0.36)	0.12 (0.34)	0.13 (0.36)	0.14 (0.37)	0.13 (0.35)		0.12 (0.35)		0.13 (0.36)
Model fit										
Model df	8	9	11	11	11	11	11	11	11	15
Test change in df	1	1	1	1	1	1	1	1	1	4
AIC	2983.9	2952.3	2942.2	2949	2905.9	2952.1	2947.6	2947.8	2951.9	2904.1
BIC	3038.8	3013.3	3009.3	3016.1	2973	3019.2	3014.8	3014.9	3019	2995.6
-2 log likelihood	-1482.9	-1466.2	-1460.1	-1463.5	-1442	-1465.1	-1462.8	-1462.9	-1465	-1437

Asterisks indicate measures that significanly predcit accuracy, * p < 0.05, ** p < 0.01, *** p < 0.001.

3.2.5.5 Accuracy characteristic plots

Following the results from the regression, we plotted accuracy characteristic plots for the measures that significantly predicted accuracy. To recap, we plotted proportion correct for younger (aged 4-6 years) and older (aged 7-8 years) groups as a function of confidence, and as a function of each significant implicit measures (e.g., open or closed-eye for box sorting; presence or absence for body and vocal measures). In all plots, the dashed line indicates chance accuracy for uncertainty indicators (e.g., at the lowest confidence, closed-eye box, slowest response times, presence of head tilt, fillers, hedges, and absence of boosters) and perfect accuracy for certainty indicators (e.g., at the highest confidence, open-eye box, fastest response times, absence of head tilt, fillers, hedges, and presence of boosters). Error bars are standard errors.

3.2.5.5.1 Confidence-accuracy characteristic

Figure 5A indicates that confidence was predictive of accuracy in both younger and older children. As in Experiment 1, confidence was more predictive for older children. Older children's high confidence responses were 94% correct, and their low confidence responses were 66% accurate. The younger children were still able to assign higher confidence judgements to their correct answers, and lower judgements to their incorrect answers; their high confidence responses were 81% accurate, and their low confidence responses 52% correct. However, their mid-point confidence responses were 81% correct, similar to their high confidence responses. This explains the significant interaction between confidence and age in the modelling and indicates that the informativeness of confidence for predicting memory accuracy improves with age; while children aged 4-6 years seem to be able to use confidence ratings to reflect their accuracy, the skill is not yet as refined as the children aged 7-8 years.

3.2.5.5.2 Box sorting-accuracy characteristic: overall

In contrast to the findings in Experiment 1, in Experiment 2, collapsed across confidence absent and confidence present conditions, box sorting was predictive of accuracy for both younger and older children. Although older children performed better generally, with open-eye answers being 88% correct closed-eye answers being 62% correct, Figure 6B shows, compared to Experiment 1, a marked improvement in the performance of younger children in sorting their answers into the appropriate box in accordance with their memory accuracy; in Experiment 2 closed-eye answers were 59% correct, and open-eye answers were 77% correct. Recall that we were also interested in if giving a confidence judgement prior to completing the box sorting task would influence children's box sorting ability. To examine this, we plotted box sorting-accuracy characteristics for the confidence absent condition (*N* = 51), and the confidence present condition (*N* = 35).

3.2.5.5.2.1 Confidence absent

For the confidence absent condition, where children provided no confidence judgement before the box sorting task, Figure 5A indicates that box sorting was predictive of accuracy for both age groups. Older children's answers in the closed-eye box were 70% correct, and 87% correct in the open-eye box. Younger children's answers in the closed-eye box were 62% correct, and 78% correct in the open-eye box.

3.2.5.5.2.2 Confidence present

Interestingly, for the confidence present condition, where children provided a confidence judgement before the box sorting task, both age groups seemed better able to assign their inaccurate answers into the closed-eye box, because in the confident present compared to absent condition, answers in the closed-eye box were lower in accuracy. Figure 5B indicates that the proportion of correct answers assigned to the open-eye box remained similar across the confidence absent and confidence present conditions. Older children's answers in the closed-eye box were 52% correct, and 89% correct in the open-eye box. Younger children's answers in the closed-eye box were 54% correct, and 75% correct in the open-eye box.

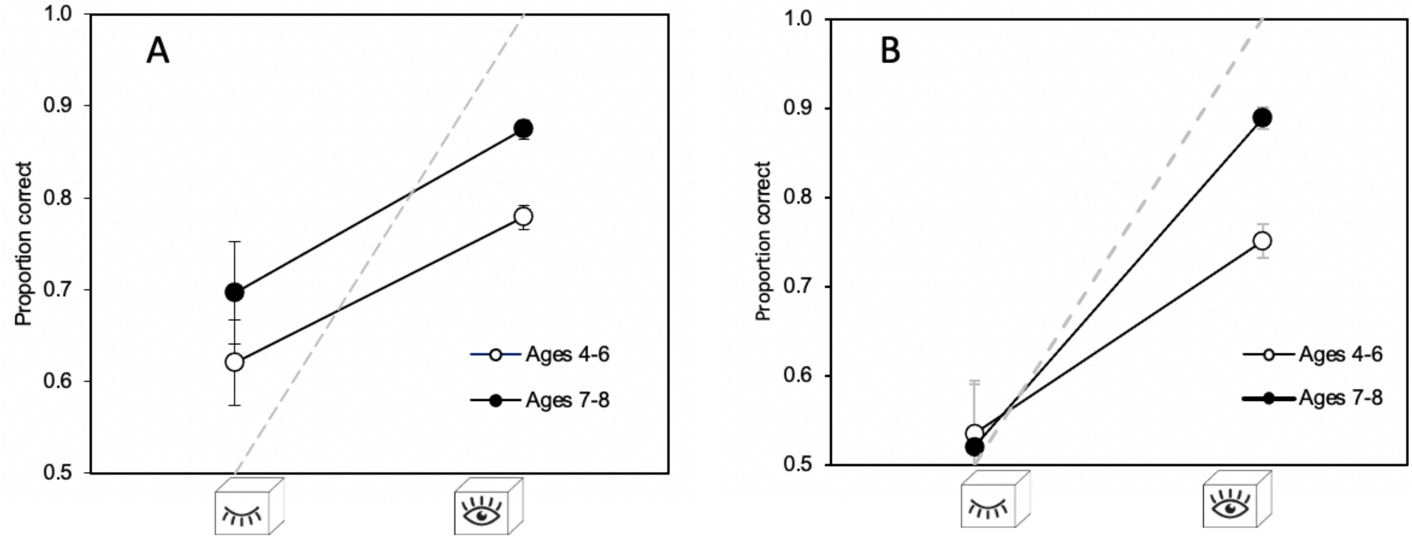


Figure 5. Accuracy characteristic plots for the confidence absent and present conditions in Experiment 2. (A) Confidence absent, (B) Confidence present. The dashed line represents chance-level performance

3.2.5.5.3 Response time-accuracy characteristic

In Experiment 1, response time was predictive of accuracy in both older and younger children. Recall that mid response times were coded as between 5 and 10 seconds (*M* = 5.72, *SD* = 1.69), with fast and slow response times falling either side of these times (Fast *M* = 2.79, *SD* = 1.03 Range = 3; Slow *M* = 16.56, *SD* = 10.46, Range = 109). We used the same response time categories for Experiment 2. Figure 6C indicates that response time was predictive of accuracy in older children and younger children, but this time in Experiment 2, it was more predictive for older children. For older children, fast responses given within 4 seconds were 92% correct, mid responses were 79% correct, and slow responses given after 10 seconds were 66% correct. For younger children, fast responses made within 4 seconds were 79% correct, mid responses were 72%, and slow responses made after 10 seconds were 65% correct. This plot can explain the significant interaction between age and response time in the regression modelling: the informativeness of response time improves with age.

3.2.5.5.4 Head shake-accuracy characteristic

Head shakes were not a predictive measure in Experiment 1. In Experiment 2, the correspondence between headshakes and accuracy was the same for both younger and older children: Figure 6D shows that when head shakes were present, children's responses were more likely to be inaccurate, although older children had higher overall accuracy. Older children were 67% correct when head shakes were present, and 86% correct when absent. Younger children were 44% correct when head shakes were present, and 75% correct when absent. However, these results should be interpreted with caution due to the low frequency of the measure indicated by the large SE bars. The regression modelling indicated that the informativeness of head shakes did not change with age.

3.2.5.5.5 Head nod-accuracy characteristic

Head nods were not a predictive measure in Experiment 1. Curiously, in Experiment 2 head nods appeared to indicate uncertainty rather than certainty. The same trend was observed in Experiment 1, although head nods were not found to be a significant predictor of accuracy. Figure 6E shows that when head nods were present, children's responses were more inaccurate. As with head shakes, older children had higher overall accuracy. Older children were 74% correct when head nods were present, and 86% correct when absent. Younger children were 44% correct when head nods were present, and 75% correct when absent. As with head shakes, these results should be interpreted with caution due to the low frequency of the measure indicated by the large SE bars. The regression modelling indicated that the informativeness of head nods did not change with age.

3.2.5.5.6 Hedges-accuracy characteristic

In Experiment 1 we found that hedges were predictive for both age groups, but perhaps slightly more predictive for younger children. In contrast to Experiment 1, Figure 6F indicates that hedges were more predictive of accuracy for older children than younger children. For older children, answers were 87% correct when hedges were absent, and 72% correct when hedges were present. The relationship was weak in the younger children, with only a 6% difference in accuracy on average between answers when hedges were absent and present; answers were 75% correct when hedges were absent and 69% correct hedges were present. The modelling indicated that the predictiveness of hedges changes with age, and this plot indicates they increase in predictiveness with age.

3.2.5.5.7 Fillers-accuracy characteristic

In Experiment 1, fillers were predictive of accuracy for both older and younger children. Figure 6G indicates that fillers were predictive of accuracy for older children only. For older children, answers were 87% correct when fillers were absent and 75% correct when fillers were present. For younger children answers were 75% correct when fillers were absent and 72% correct when fillers were present. Older children achieved a higher proportion of correct answers when fillers were absent than younger children. The plot and regression modelling suggest that fillers were predictive of accuracy for older children only, and therefore that predictiveness increases with age.

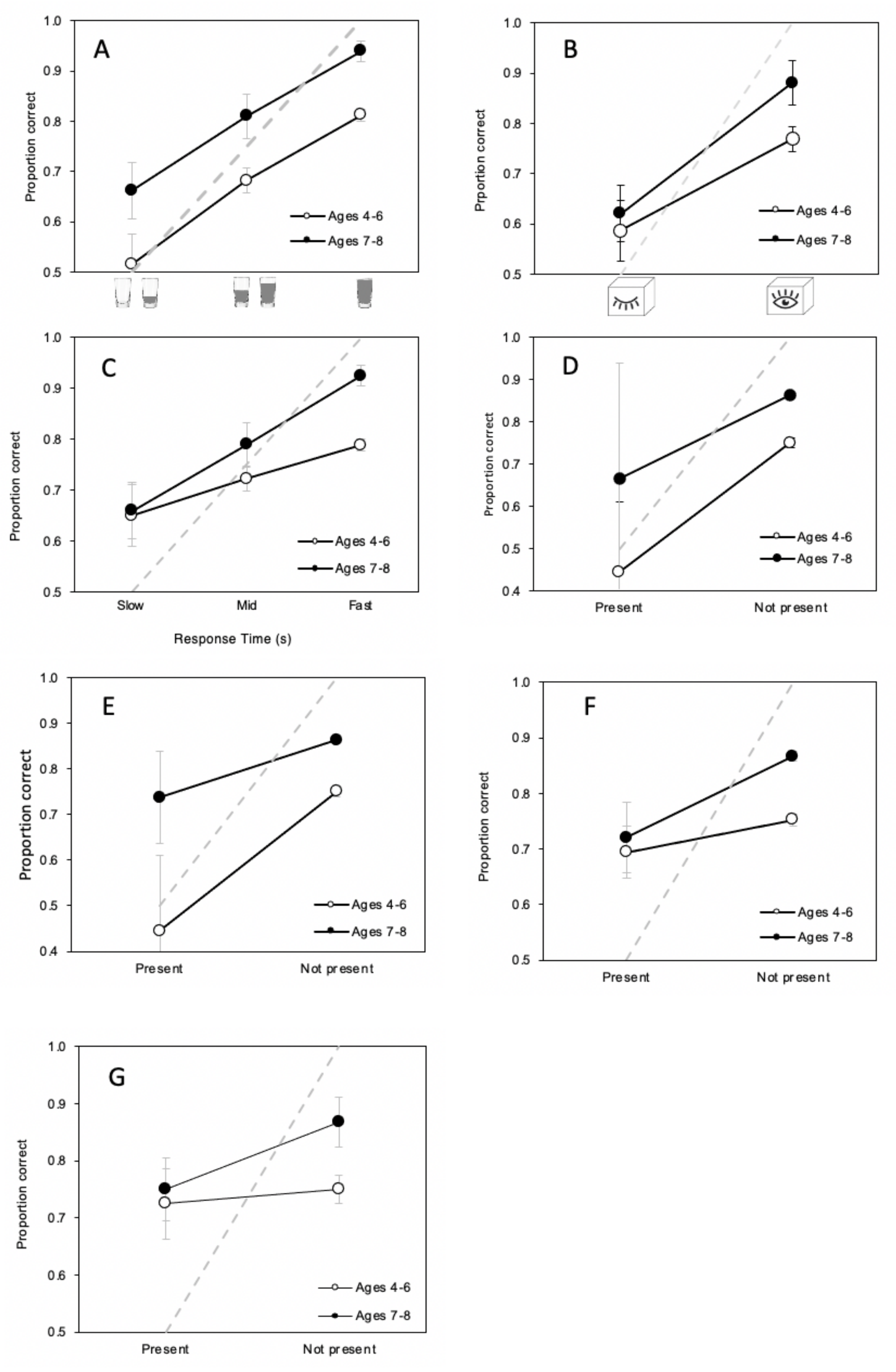


Figure 6. Accuracy characteristic plots for the 7 significant measures in Experiment 2. (A) Confidence, (B) Box sorting, (C) Response time, (D) Head shakes, (E) Head nods, (F) Hedges, (G) Fillers. The dashed line represents chance-level performance.

3.3 Summary

In sum, the results from Experiment 2 indicate that confidence, box sorting, response time, head shakes, head nods, hedges and fillers were predictive of accuracy. Confidence, box sorting, response time and age were uniquely predictive of accuracy when controlling for other measures, with confidence being the most predictive measure and more predictive of accuracy than age. Contrary to Experiment 1, response time remained a unique predictor when age was added as a main effect. Finally, the informativeness of confidence, box sorting, response time, hedges and fillers changed with age: they all became more informative with age. The change in informativeness of response time with age was the most significant.

The results across both Experiments are discussed and linked to the broader literature in Chapter 4, next.

CHAPTER 4

GENERAL DISCUSSION

4.1 Summary

The aim of this thesis was to connect the eyewitness and developmental literature, and to apply the developmental theory of metacognition to resolve a forensic problem. Specifically, the thesis investigated which implicit and explicit metacognitive measures best predicted accuracy in children aged between 4 and 8 years after encoding a complex episodic event. We were also interested in the development of uncertainty monitoring and control in children of this age range, and the relationship between the two processes. This final chapter will summarise the findings from the two Experiments and discuss the results considering previous research and address limitations and possible future directions.

4.2 Findings

4.2.1 Memory performance

True to literature on developmental improvement in memory task performance (e.g., Bauer & Fivush, 2014), and retrieval processes (e.g., Roebers C. M., 2013), older children performed significantly better on the memory task than younger children (as demonstrated in the higher overall accuracy for both experiments). As such, age differences found in monitoring and control ability could be attributed to task difficulty, with older children excelling in both due to having better underlying memory performance (Winsor et al., in press). Nevertheless, in forensic contexts memory performance is likely to be different for younger and older children and therefore it is important to determine which metacognitive measures predict accuracy in different ages.

4.2.2 Measures that were predictive of accuracy

Overall, the results from Experiment 1 indicate that predictors of accuracy in children aged 4-8 years include confidence, box sorting, response time, fillers, hedges, head tilt, and boosters. Shrugs, looking to caregiver, thinking gesture, head nods and head shakes did not predict accuracy. The results from Experiment 2 partially replicated these findings, with confidence, box sorting, response time, fillers, hedges predicting accuracy alongside the novel results of head nods, and head shakes. Shrugs, looking to caregiver, thinking gesture, head tilts and boosters did not predict accuracy. Here, we can conclude that the most consistent measures that predicted accuracy were confidence, box sorting, response time, hedges and fillers, as these measures were significant across both experiments. Whilst head tilts, head nods, head shakes and boosters were interesting findings, they did not replicate across both experiments, indicating that they cannot be considered stable predicators of accuracy.

4.2.3 Measures that uniquely predicted accuracy

To determine which measures uniquely predicted accuracy, we z scored our 7 significant measures. Unique predictors of accuracy for Experiment 1 included confidence, response time and head tilts. Confidence was the most uniquely predictive, followed by response time and head tilts respectively. These results suggest confidence offered the most information about children's accuracy when controlling for other measures, followed by response time and head tilts. When age was added as a main effect in Experiment 1, confidence, head tilt and age were uniquely predictive of accuracy, but response time was no longer a predictive measure. As such, it could be that the unique portion of information explained by response time was explained by age when added to the model, meaning that response time no longer offered unique information.

Unique predictors of accuracy for Experiment 2 included confidence, response time and box sorting. As in Experiment 1, confidence was the most uniquely predictive, followed by response time and box sorting, respectively. These results suggest confidence offered the most information about children's accuracy when controlling for other measures, followed by response time and box sorting. When age was added as a main effect in Experiment 1, confidence, response time, box sorting and age were uniquely predictive of accuracy.

Confidence and response time were consistent unique predictors across both Experiments, but *only when the main effect of age was not considered*. As such, when age was considered, *confidence was the most stable unique predictor of accuracy* as it was consistently a unique predictor across Experiments. In sum, for both Experiments, whilst the significant measures were useful in predicting accuracy, confidence provided the most unique information. In other words, confidence explains a unique portion of the variance in accuracy, and so theoretically, it is possible that the non-unique predictors for both experiments could be considered as proxies for confidence. For example, a filler may be a behavioural indication of low confidence, and a booster an indication of high confidence.

4.2.4 Measures that were more predictive of accuracy than age

To determine which measures were more predictive of accuracy than age, we z scored our measures and added age as a main effect to the regression model alongside the 7 significant predictors. The results indicated that confidence was the only predictor *more* predictive of accuracy than age in both Experiments. This is an integral finding for two reasons: first, it indicates that confidence offers *more information* about children's likely memory accuracy than their age. Therefore, relying on age to determine memory accuracy (e.g., Knutsson & Allwood, 2014; Newcombe & Bransgrove, 2007; Wigmore, 1935/1976) may not be the best method, as confidence was a more informative measure. Second, it suggests that explicit measures are *more informative of memory accuracy* than age and implicit measures in younger children. This suggest that younger children's explicit metacognitive abilities may be more advanced than previously believed. It also may indicate that confidence is more useful than implicit measures because children were required to make a confidence response on each trial, but did not always use implicit gestures or vocal measures to indicate their certainty.

4.2.5 Measures that changed with age

To examine if any of the measures changed in informativeness with age, we added age as an interaction effect to the regression model. In Experiment 1, adding age as an interaction effect and plotting accuracy characteristic plots indicated that the informativeness of confidence, head tilt and box sorting changed with age. In Experiment 2, adding age as an interaction effect and plotting accuracy characteristic plots indicated that the informativeness of confidence, response time, box sorting, hedges and fillers changed with age. We further explore these findings in the following section.

Whilst confidence was predictive of accuracy for both age groups in Experiment 1 and 2, it was more predictive of accuracy for children aged 7-8 years in both Experiments. This is also consistent with developmental literature that states that children as young as 4 can make accurate confidence judgements (Hembacher & Ghetti, 2014), but the ability improves with age. Younger children have been deemed to be 'eternal optimists' (Mickes, Hwe, Wais, & Wixted, 2011), and it is clear from the confidence-accuracy plot, that younger children were more confident than they ought to have been considering their level of accuracy at high confidence (i.e., selected the full cup confidence when they were 84% accurate in Experiment 1 and 81% accurate in Experiment 2, not 100% accurate). Younger children also more frequently used the highest confidence rating (full cup). In Experiment 1, younger children assigned the highest level of confidence to 68% of their answers, whilst older children assigned the same confidence rating to only 44% of their answers.

Likewise, in Experiment 2 younger children assigned the highest level of confidence to 67% of their answers to the highest level of confidence, whilst older children assigned the same confidence rating to only 51% of their answers. These results are consistent with the notion that uncertainty monitoring, and therefore the ability to explicitly assign appropriate confidence judgements that reflect accuracy, improves with age (e.g., Hembacher & Ghetti, 2014; Winsor et al., in press). Kloo & Rohwer (2012) note that children under the age of 6 years tend to be overconfident in their estimations of accuracy as they are not yet competent in explicitly indicate their certainty. Whilst our results show that younger children do seem able to explicitly indicate their certainty, Kloo and Rohwer's explanation for younger children's overconfidence accounts for the improvement in older children's confidence-accuracy relationship. It is also consistent with previous research that has found that younger children's level of confidence tends to be higher than their level of performance (Finn & Metcalfe, 2014; Roebers, Kälin, & Aeschlimann, 2019; Winsor, et al., in press).

Despite the overconfidence at high confidence in the younger children, the results do indicate that confidence ratings in children as young as age 4 years are indicative of accuracy after encoding a complex episodic event. This conclusion is at odds with previous eyewitness literature, which has consistently stated that children under the age of 12 years cannot give accurate confidence judgments in a forensic context, but consistent with the developmental literature which has used list-learning memory studies (e.g., Hembacher & Ghetti, 2014). Together, the use of a complex episodic event as a task for these experiments, and consideration of the broader literature indicates that children appear to be able to provide appropriate explicit confidence judgements for both simple tasks and the complex memory task we used here (see also Winsor et al., in press).

In their Experiment testing children aged 6-13 years on an eyewitness identification task, Bruer and colleagues (2017) highlighted a concern that the children would use the confidence scale dichotomously by only choosing between the 2 extremes (empty cup and full cup). This trend has often been observed in younger children (e.g., Goodenough, et al., 1997; von Baeyer & Webb, 1997). Younger children did seem to use the middle of the scale less frequently that older children. In Experiment 1, younger children assigned only 23% of their answers to the middle of the scale, whereas older children assigned 40% of their answers to the middle of the scale. This was also true for Experiment 2, where younger children assigning only 21% of their answers to the middle of the scale, whereas older children assigned 38% of their answers. This again suggests that children become better accustomed to using confidence scales appropriately with age, perhaps because older children are more able to make finer discriminations between their feelings of certainty, but also perhaps because older children have a better understanding about how to use scales with more that 2 points (Chambers & Johnston, 2002).

Contrary to Hembacher and Ghetti's (2014) findings, in Experiment 1, we found that children aged 4-6 years did not have a relationship between box sorting and accuracy. 7–8-year-olds had a good box sorting-accuracy relationship, indicating older children were able to successfully sort their answers based on accuracy. Similarly, in a sorting task, children were asked to sort animals into appropriate boxes, only children aged 6 years and older could appropriately use an empty box to avoid making an error (Kloo & Rohwer, 2012b). In Experiment 1, younger children appeared to understand the instructions in the practice phase but then did not use the boxes appropriately. Children may be able to verbally represent a rule but are unable to apply that knowledge to guide their behaviour on task. It could be the case that the children under the age of 6 years were able to understand the box-sorting instructions on a conceptual and linguistic level but were unable to act strategically or employ self-regulation on the box-sorting task (Lyons, 2011; Russell, 1997). This idea is also consistent with the production deficiency hypothesis, as children may identify that their memory is poor, but not know how to utilise the tools provided to indicate this appropriately when choosing what to report.

Conversely, the results from Experiment 2 painted a more positive picture of the younger children's control abilities. Although older children were again more adept at the box-sorting task, younger children also showed a relationship between box sorting and accuracy and seemed able to generally sort their correct answers into the open-eye box, and their incorrect answers into the closed-eye box (see Figure 6B). This is in line with results from Frye (1995), who found that children over the age of 4 years were able to strategically control their actions.

What could explain the younger children's poorer performance on the box sorting task in Experiment 1 compared to in Experiment 2 and observed by Hembacher and Ghetti (2014)? One possible explanation of poorer box sorting performance in Experiment 1 compared to Hembacher and Ghetti could be task difficulty. Bryce & Whitebread, (2012) found monitoring and control processes improve with age, but that the deficits in metacognition are primarily affected by the task involved rather than age. Younger children could have failed to use the boxes correctly to indicate accuracy because we used a complex to-be-remembered event, whereas Hembacher and Ghetti (2014) used a simple memory task that involved remembering line drawings. Some might argue that complex to-be-remembered events are cognitively more demanding than basic list learning paradigms. If a child finds the task difficult, metacognitive control may be interrupted and hindered by task-related anxiety (Zeidner, 1998). Task irrelevant thoughts, such as preoccupation about poor task performance may distract the child from implementing metacognitive control (Kanfer & Ackerman, 1996). Similarly, tasks that involved a to-be-remembered event may require more cognitive resources, leaving little room for metacognitive control (Fox, Park, & Lang, 2007). This explanation, however, cannot account for the finding that box-sorting was informative of accuracy in younger children in Experiment 2, as the same task was used across Experiment 1 and 2.

Instead, the variability in younger children's box sorting ability across Experiments could be due more children completing all the trials in Experiment 2. More younger children in Experiment 1 chose to stop after 20 or fewer trials. It could be that it took younger children longer to understand how to use the boxes, and so their ability to use it correctly was maximised when all 40 trials were completed. That is, perhaps they did not complete enough trials in Experiment 1 to understand the task sufficiently. Alternatively, perhaps the younger children recruited in Experiment 1 were just less able or motivated to complete the task than the children recruited in Experiment 2. Whatever the reason for the conflicting results, conclusions should be drawn with caution and further research should explore children's box sorting (i.e., metacognitive control with verbal report) across different child samples and tasks.

The regression analyses revealed that for Experiment 1 the informativeness of response time did not change significantly with age. Figure 3C indicates a relationship between response time and accuracy for both younger and older children: all ages responded more slowly when inaccurate, and more quickly when accurate, although the plot does indicate that 7-to-8-year-olds children had a higher proportion of correct answers when responding within 4 seconds than 4-to-6-year-olds (see Appendix F). However, Experiment 2 indicated that response time became more informative with age, with it being a better indicator of accuracy in older children. This mirrors previous results that have found age related increases in the predictiveness of response latency (Ackerman & Koriat, 2011; Koriat & Ackerman, 2010a). Children as young as 25 months have been shown to look longer at items when they are unsure (e.g., Leckey, et al., 2020), and so it follows logically that response time would continue to indicate uncertainty in children beyond this age. Indeed, in line with previous research (e.g., Koriat & Ackerman, 2010; Ackerman & Koriat, 2011; Lyons & Ghetti, 2011) response time was predictive of accuracy for both younger and older children.

Considering hedges, the initial conclusion from the first Experiment was they were predictive of accuracy for both age groups, but younger children had a greater difference in proportion correct between presence and absence of hedges. This finding can be considered drawing from Vygotsky's theory of private speech (Vygotsky, 1962; Winsler, Fernyhough, & Montero). Namely, younger children appear to engage more in audible utterances than older children when performing tasks. The data reflected this, as although the frequency of hedges performed was relatively similar across age groups, it seemed that younger children were more likely to be inaccurate when performing them than older children. However, the results from Experiment 2 suggested a different conclusion. Hedges changed in informativeness with age, and were more predictive of accuracy in older children, and younger children had a weaker relationship. The number of trials for younger children and older children were more equal in Experiment 2, and so we could infer that the Experiment 2 results are more representative of the informative of hedges.

Fillers were predictive for both age groups in Experiment 1, but predictive only for older children in Experiment 2; younger children had no relationship between accuracy and fillers. Some studies suggest that the use of fillers in speech increases with age (e.g., Esposito, Marinaro, & Palombo, 2004; MacWhinney & Osser, 1977). Whilst this notion is supported by the findings in Experiment 2, the results for Experiment 1 suggest that both older and younger children performed fillers when uncertain. Although the frequency of fillers performed was relatively similar across age groups, it could be the case that only older children performed them when inaccurate, and that there was no predictive pattern with the younger children's performance of them. As with hedges, the number of trials for younger children and older children were more equal in Experiment 2, suggesting that the Experiment 2 results may be more representative of the informative of fillers.

In Experiment 1, head tilts were more predictive of accuracy in older children than in younger children. This change in informativeness with age could be attributed to the fact that certain behaviours are socially learnt, and that the younger children have not learnt them from observed behaviours yet. Piaget (1950) asserts that children's awareness of social appearance increases with age, suggesting that social settings are integral in the development and display of gestures of certainty and uncertainty. For example, one study found that children begin using head nods and head shakes at around 8 months after having observed adults ((Fusaro, Vallotton, & Harris, 2014; Goldin-Meadow, 2015). It is possible that this learning-by-observation could be the basis for head tilts too. Similarly, younger children may not have observed others using this measure in this specific context, and so do not yet associate the gesture with uncertainty (Krahmer & Swerts, 2005). However, head tilts were not a predictive measure in Experiment 2, and so conclusions concerning head tilts should be interpreted cautiously, as it was not a stable predictor across experiments.

4.2.6 Measures that did not change with age

Boosters were predictive for both age groups in Experiment 1, but not Experiment 2. In Experiment 1 both age groups had a good relationship between accuracy and boosters, but older children had higher overall accuracy than younger children when boosters were present and when booster were absent. Initially it appears that this result contradicts Vygotsky's theory of progressively internalised speech (1962). However, if we look at the frequencies (see Appendix F and Appendix H), it appears that younger children performed a higher number of boosters when correct than older children, but older children had a higher rate of accuracy when performing them. This suggests that older children's boosters are more predictive than younger children's: although older children do not use boosters as frequently as younger children, when they do, they are more likely to be accurate than younger children. Another interesting theory described by Vincze & Poggi, (2016) is that children may use boosters when unsure to avoid subsequent questioning. This may account for the higher frequency yet lower predictiveness of booster in younger children: they may use boosters even when unsure to avoid being questioned further about something they are unsure about (Vincze & Poggi, 2016). Since boosters were not a significant measure in Experiment 2, they cannot be considered a stable indicator of accuracy.

Head nods were a significant predictor of accuracy for both age groups in Experiment 2, but not for Experiment 1. Previous research has concluded that head nods are a stable gesture of certainty in both adults (e.g., Borràs-Comes, Roseano, Bosch, Chen, & Prieto, 2011) and children (e.g., Fusaro & Harris, 2013; Harris, Bartz, & Rowe, 2017), yet contrary to that, our results from Experiment 2 suggested that head nods were performed more frequently when children felt uncertain rather than certain (and the same trend was observed in Experiment 1). Although some research has found that head nods can convey negative evaluations, these negative evaluations have still been in the context of something being understood rather than not understood (e.g., a message being understood and rejected; Cowie 2002). Given the host of research on head nods as a certainty gesture, our finding that head nods indicated uncertainty in Experiment 2 is an odd result which could be explained by children of this age range in our sample not having socially learnt that head nods can be used as a marker of certainty. This could also be the reason for the low frequency and lack of predictive value of head nods in Experiment 1. Considering both experiments and the previous literature, our results suggest that, at least on that task and coding scheme that we used here, head nods do not convey consistent or useful information about likely memory accuracy in children aged 4-8.

Similarly, head shakes were a significant predictor of accuracy for Experiment 2, but not for Experiment 1. The findings from Experiment 2 are consistent with previous research that head shakes are indicators of uncertainty (e.g., Harris, Bartz, & Rowe, 2017; Kendon, 2003). The frequency of headshakes was fairly low in Experiment 2 (see Appendix G) and even lower in Experiment 1 (see Appendix E), owing to the smaller sample size. Although head shakes were performed infrequently, they were almost always performed when a child's answer was inaccurate, therefore signally uncertainty. As Experiment 2 had a larger sample size than Experiment 1, it could be that Experiment 1 had too small a sample size to observe this effect. If sample size was indeed the reason for the difference in findings, then it could be plausible to conclude that headshakes are a stable predictor of uncertainty in children for a complex memory event. This would be in line with the current robust body of evidence that exists for headshakes as an uncertainty gesture. However, further research with larger samples is required to confirm that conclusion.

4.2.7 Measures that did not predict accuracy

Looking to caregiver did not predict accuracy in either Experiments. Although children appeared to frequently use this gesture (see Appendix E and Appendix G), results from the regression indicated that it was one of the least informative measures of accuracy (see Table 4). This could be because children often performed the action, but not specifically during trials where their answers were incorrect. This suggests that the gesture may not be related to uncertainty, at least on the task that we used here. Previous research has found that children do look to a caregiver when they are uncertain (e.g., Campos & Steinberg, 1981). Prior to our experiment, adults were asked not to help their child during the task. Children were present during these instructions, and often parents would reiterate this to them (e.g., telling the child, "I'm not allowed to help you, okay?"). Children may therefore have been aware that looking to the adult for information or feedback would not be a useful strategy during the task if they were uncertain. If this was the case, why was the frequency of the gesture so high, if not for information seeking? Children's proximity seeking behaviour has been found to be an 'inborn affect-regulation device' that aids in the alleviation of distress (Bowlby, 1982; Conner, et al., 2012). The circumstance of this task (e.g., being tested by a stranger) may have been anxiety inducing for the children, and so looking to caregiver may have been an act of seeking assurance, rather than task related information or feedback.

Shrugs were the least occurring measure in both experiments, and therefore not a useful indicator of memory accuracy on this task. It is perhaps unsurprising that we observed few instances of shrugs when children were uncertain in our study, as the literature suggests that shrugs can convey different meanings. For example, Kendon, (2004) asserts that shrugs fall into the semantic domain of ‘interpersonal control’ and may be utilised to an individual’s interpersonal attitude. Further qualitative research has suggested that, in native English speakers, shrugs convey disengagement and disinterest (Streeck, 2009). They can also indicate obviousness (Debras, 2017; Vincze & Poggi, 2016). It could be that this task did not provide the context for the performance of shrugs. Nevertheless, it could be considered a positive outcome that children performed few shrugs, as the previous literature would suggest that—due to the low occurrence—children were engaged and interested in the experimental task.

Overall, the presence of explicit and implicit metacognitive measures and their relationship with accuracy indicates that children as young as 4 can monitor their uncertainty and appear to be able to express this explicitly and implicitly.

4.2.8 Confidence and box sorting relationship

Considering the box sorting and confidence results together, in Experiment 1, younger children were able to use confidence judgements, but unable to use the box sorting decision to indicate their accuracy. This is in line with the developmental trajectory observed in monitoring and control abilities. Younger children may be able to monitor their uncertainty, but do not change their behaviour accordingly. Other research has found that monitoring processes are present in 6-year-olds, but that control processes are yet to mature (Bryce & Whitebread, 2012).

One of the aims of Experiment 2 was to further explore how children performed on the box sorting task when confidence was not a precursor. We wanted to further investigate the relationship between monitoring and control in children of this age group. As discussed previously, Experiment 2 yielded different results; both age groups seem to successfully reflect on their monitoring when confidence was present, as the proportion correct for the closed-eye box was lower than when confidence was absent. The novel finding in Experiment 2 was that children in both age groups who were asked to rate their confidence *before* performing the box sorting task showed a *stronger relationship* between box sorting decision and accuracy.

Interestingly, this trend appeared to be true only for inaccurate answers; proportion correct for the accurate answers remained similar for both age groups. It could be that having the confidence judgements before the box sorting task negated overconfidence in inaccurate answers for all age groups. This finding is in line with previous research that suggests confidence judgements facilitate the discrimination between accurate and inaccurate responses and subsequent strategic behaviour (in this case, box sorting) (e.g., Hembacher & Ghetti, 2014). The relationship between box sorting and accuracy existed for children in the confidence absent condition too, although the proportion correct for answers sorted in the closed-eye box was higher than for the confidence present condition. These results suggest that children’s confidence judgements may indeed influence their later strategic behaviour. It appears that confidence better informed the children’s closed-eye box sorting decision, and by asking children to rate their confidence first they thought more about whether they wanted their answer to be hidden. If this is the case, then it provides interesting information for application. For example, asking a child how sure they are of an answer, then asking them if they would like the answer to be recorded or seen may provide legal decision makers with a more accurate representation of when children are inaccurate, than if the child is not asked how sure they are.

Previous research (e.g., Hembacher & Ghetti, 2014) collected confidence then box-sorting decisions and concluded that box-sorting is informative of accuracy in young children from age 4. However, this conclusion may be premature if box-sorting proficiency (i.e., metacognitive control) is dependent on providing a confidence judgement (i.e., metacognitive monitoring); it could be the case that children can only utilise the box sorting task when it is preceded by confidence.

From the results across experiments, it appears that older children have superior monitoring and control skills. In terms of monitoring, older children had a strong confidence-accuracy relationship (see Figure 3A & Figure 6A) and were able to appropriately rate their confidence in accordance with accuracy. In terms of control, older children performed well in the box sorting task across both experiments. They were able to both sort their answers into the appropriate box according to accuracy (see Figures 3B & Figure 6B) and confidence rating. This is unsurprising, as there is a host of research to suggest that monitoring and control skills improve with age (e.g., Bryce & Whitebread, 2012; Hembacher & Ghetti, 2014; Lyons, 2011; Schneider, 2000). These results are in line with this evidence that between the ages of 4 and 8 years are a developmentally sensitive period. As suggested by Vygotsky, older children may be more adept at monitoring and controlling their behaviour as these skills are more likely encountered socially and enhanced in school (e.g., Lockl & Scheider, 2002; Roebers, Schmid, & Roderer, 2009).

4.3 Variability among children

The differing result across experiments suggests variability across different samples of children. Indeed, between 4-8 years old is considered to be a key developmental period for metacognition, and as such children's metacognitive ability may progress at different rates

(Flavell, Beach, & Chinsky, 1966; Keeney, Cannizzo, & Flavell, 1967; Veenam, Van Hout-Wolters, & Afflerbach, 2006). Children with higher IQs have been shown to have better metacognition (Calero, Garcia-Martin, Jiménez, Kazén, & Araque, 2007; Swanson, Kehler, & Jerman, 2010; Shore & Dover, 1987). Importantly, children may also have different experiences and opportunities that allow them to develop their metacognitive ability; for example, some schools may employ more activities for building metacognitive ability than others. Children may therefore function at different levels of metacognition, resulting in greater variability between children. The role of relations in school in the development of metacognition is highlighted by Vygotsky's research (1962). The results of these experiments may be testimony to the importance of consistent skill teaching in schools to ensure all children develop sound metacognition. In terms of implicit body and vocal measures, collecting data from a larger sample size of children may be useful in determining which measures are stable predictors of accuracy in children across samples, and which are not. Therefore, the results may be variable across experiments due to different children being sampled across experiments.

4.4 Practical Implications

From this discussion, we can conclude that stable predictors of accuracy on our experimental task for children aged 4-8 years appear to be confidence, response time, box sorting, hedges, and fillers. These measures were significant predictors of accuracy across both experiments. When considering age in the model, confidence was consistently a unique predictor of accuracy over both experiments, with response time and box sorting as unique predictors for Experiment 2. In practice, the finding that confidence is a unique predictor of accuracy means that collecting an implicit measure (e.g., hedges) alongside confidence would provide no additional information about accuracy than if just confidence was collected. Confidence was the most informative measure of memory accuracy across both experiments and was more informative that age. Therefore, the most informative way to collect information about the likely accuracy of memory statements made by child witnesses aged 4-8 years appears to be to ask them for a confidence judgement. Indeed, the adult memory literature has also recommended that confidence judgments be collected from adult witnesses (e.g., Wixted & Wells, 2017).

Response time was a consistent significant predictor across experiments, and also provided unique information in addition to confidence in Experiment 2. Response time being a useful indicator of accuracy has also been replicated in the adult literature (Seale-Carlisle, et al., 2019). Interviewers may also wish to probe for more information when interviewees take a long time to respond to questions or note down which answers witness were slow to respond to. Box sorting was also a strong predictor of accuracy and offered unique information in Experiment 2. In practice, box sorting could be also used by asking a child witness if they would like their answer to be seen or not be seen by another interviewer; or possibly by informing children that it is fine to withhold answers for which they are unsure (e.g., see Lyons & Ghetti, 2013). However, conclusions regarding response time and box sorting should be drawn with caution as there were inconsistencies across experiments. Future research could further explore the use of the box sorting task to better understand how useful it is in predicting accuracy for children of this age.

Yet, it is still important to consider that some critics might highlight that confidence, response time, and box-sorting decisions may not be considered currently to be the most convenient measures to collect in practice. For example, one concern might be that collecting confidence ratings or box-sorting decisions may undermine the flow of an interview, especially a witness's free-recall account, if interviewers are required to stop and ask for certainty judgements after each piece of information provided by the witness (Fisher & Geiselman, 2010). Moreover, it may be difficult to accurately measure the speed of an individual's response during an interview because we know that the cognitive load on the interviewer is already high (Hanway, Akehurst, Vernham, & Hope, 2020).

However, research suggests that these possible issues are of limited concern, at least when memories are collected from adult witnesses. For example, Spearing & Wade (2021) found that the confidence-accuracy relationship was as strong in adults when the confidence ratings were collected after the interviewer had asked all the questions, compared to when the interviewer stopped after each question to obtain a confidence judgement. Additionally, one field study had investigators collect response time information from real adult witnesses making an identification decision from a lineup (Wells, 2014) and response time predicted accuracy (Seale-Carlisle, et al., 2019). Results like these are promising for the collection of confidence and response time in a legal setting. Such studies are yet to be conducted with child witnesses, however. Therefore, while fillers and hedges provide no information over and above unique predictors (confidence, response time, and box sorting), they may be considered practical measurements to collect in the context of a police interview, as they may be more easily noted when they occur. For practical application, future work should jointly consider the relative informativeness of each measure, while also considering which measures might be able to be collected and used in practice.

Another important practical point of note is that our findings apply to the memories and certainty reported at initial recall or questioning. It is well known that confidence ratings and memory reports from both adults and children can be manipulated by external information (e.g., Kassin & Kiechel, 1996; Loftus, 1994). Children are believed to be highly suggestible and vulnerable to misinformation, both in turn negatively impacting their memory recall (Holliday, Brainerd, Reyna, & Humphries, 2009). Leading questions and challenging assumptions during interview have been shown to reduce the confidence-accuracy relationship in children (Kebbell & Johnson, 2000). This information may deter police from wanting to collect children's certainty judgements as they may have been influenced by feedback or co-witness information (Benton, Ross, Bradshaw, Thomas, & Bradshaw, 2006). Nevertheless, it appears that, just as in the adult witness literature, children's expressions of certainty at initial recall are informative about memory accuracy.

4.5 Limitations & Further Research

Some limitations were encountered during the experiments. First, due to the experiments being conducted online, it was difficult keeping the set up consistent across participants as it varied depending on whether the participants were using a laptop or a desktop computer. Those with laptops had more freedom with webcam movement, and so could be asked to alter the angle if the child could not be seen. Desktop computers with built in webcam were difficult to navigate, as they could not be altered as easily. This also made seeing when the children were pointing difficult. Many children, particularly the younger ones, moved around a lot during the experiment, making it difficult to keep them within the video frame. This is an unfortunate caveat of an online experiment. Future online research could benefit from considering these issues.

Moreover, because online studies with children have only recently become more popular due to COVID-19 restrictions, it was difficult to make informed decisions about some elements of the methodology. For example, it was difficult to decide whether to have the experimenter's camera on or off during the experiments. On the one hand, a child's engagement may depend on whether the experimenter can be seen on Zoom. Studies with children have shown an increase in performance when the experimenter was present, and that experimenter presence can promote task engagement (e.g., Draeger, Prior, & Sanson, 1986). On the other hand, children may have been attending more to experimenter video presence than to the task, possibly affecting their performance. Experimenter evaluation from experimenters who are viewed as experts can also cause increased apprehension in participants (Cottrell, Wack, Sekerak, & Rittle, 1968). Participant engagement and how the experimenter could potentially impact this should be carefully considered in future research conducted online with children.

Research also suggests an experimenter who leaves room causes increased arousal from anticipation of evaluation paired with being unable to monitor the experimenter's behaviour (Guerin, 1986). It could be the case that not having a visual of the experimenter but hearing their voice intermittently could make their presence unpredictable and so caused some children to be nervous. Again, future research could consider how their presence will affect participant's engagement and demeanour.

Encouragement was given to the children throughout the experiment to motivate them to finish the study. We were careful not to give encouragement exclusively after correct or incorrect answers, but rather space it through the experiment. Similarly, feedback was given to the children after the practice phase. High levels of overconfidence have been observed in children after receiving positive feedback (Allwood et al, 2005b). Future research may benefit from not giving encouragement during the task to avoid this effect, although it would be hard to avoid giving confirmatory feedback for understanding instructions.

Another limitation of the current studies is that we used an image-based 2AFC memory task, which is far removed from the way in which memories from children are collected in the CJS. Conducting these experiments was an important first step in establishing if metacognitive measures are informative in children aged between 4-8 years. However, children in our task may use strategies that they are not able to use in real police interviews. For example, our task meant that children could verbally express that they did not know the answer to a question, but they would still have to pick an answer to proceed. Similarly, and perhaps because of this, there were instances where children appeared to be using a recall-to-reject strategy (e.g., "it can't be that one, so it must be this one" or "I don't remember seeing that item, so it must be the other item"). Schmid and colleagues (2010) noted that recall-to-reject decisions can result in high confidence rejections. It could be the case children are making their decisions by the recognition of negative evidence and so their confidence judgments may reflect their certainty that the item they have rejected was not present in the video, rather than their certainty that the item they chose was in the video. Therefore, to continue to build on this research and to make it more applicable for practice, future research would now benefit from testing children using cued-recall or free-recall testing conditions that are more akin to the way in which memory is collected (and therefore possibly more realistic of the strategies used by witnesses) in the CJS. Similarly, in a real life case, delays of various lengths are likely before a child witness is asked to recall and provide memory evidence (Humphries, Holliday, & Flowe, 2012).

Our experiments do not consider individual differences that may exist across children. For example, highly anxious individuals demonstrate lower levels of metacognitive ability (Veenman, Kerseboom, & Imthorn, 2000; Veenam, Van Hout-Wolters, & Afflerbach, 2006). Similarly, some children became shy once the experiment began.. Shyness has been found to be related to a lower confidence level (Kleitman & Stankov, 2007). At the beginning of the experiment, we would remind the adult to not help the child. At this reminder, many children would express anxiety at not being able to ask for help. Task anxiety may result metacognitive deficiency, as cognitive capacity is being absorbed through self-preoccupation (Veenman, Kerseboom , & Imthorn, 2000). Both may hinder the child's metacognition. This is especially true if the abilities are not yet fully ingrained, perhaps as is the case in younger children (Veenman, Kerseboom , & Imthorn, 2000; Veenman, Van Hout-Wolters, & Afflerbach, 2006). This could be the case for anxious children: they may have been unable to utilise their metacognitive skills because they were too preoccupied about their performance. Similarly, as the task was relatively long, task irrelevant thoughts may have added to the cognitive load during the task, hindering the utilisation metacognitive skills (Everson, Smodlaka, & Tobias, 1994).

Similarly, there is a host of research that indicates children who experience adverse childhood experiences (ACEs) and traumatic events in early life have impaired cognitive development and cognitive deficiencies (e.g., Bücker, et al., 2012; Gould, et al., 2012; Jaffee & Maikovich-Fong, 2010; Nemeroff, 2004). It would be interesting and relevant to legal settings for future research to explore metacognitive deficits specifically in children with ACEs and compare these findings to neurotypical children.

Finally, future research could explore how the confidence-accuracy relationship is affected by different types of confidence scales.

4.6 Conclusion

This thesis considers the basic and applied literatures and presents two experiments on metacognitive ability and development in neurotypical children, considering the potential forensic applications. The novel elements of this thesis were considering possible measures of metacognition (from the basic developmental literature) but using a complex episodic event for encoding mimicking conditions experienced when a crime is witnessed and employing accuracy characteristic analysis for the results (from the applied eyewitness literature). As discussed in the thesis, previous literature has not yet considered both basic and applied methods in this way to test the metacognitive abilities of children aged between 4-8 years on a complex episodic memory task. Across both experiments, our results suggest that children from the age of 4 years generally have good metacognitive ability and are aware of when their memories are not accurate. The changes in informativeness of measures between the ages of 4-8 years further demonstrates the developmental trajectory of metacognition. Confidence and box sorting ability were shown to increase in informativeness with age. This is stable evidence to infer that metacognitive ability develops within this age range. Our results also show that some metacognitive measures–such as confidence–provide unique information useful for assessing the accuracy of children's statements about a complex episodic event. Moreover, hedges also consistently predicted accuracy for all ages across the two experiments, and fillers appears to be a stable predictor for older children.

Perhaps the most interesting finding across both experiments was that confidence was consistently predictive for all age groups across conditions. This is in line with developmental literature, and more research eyewitness literature suggesting that confidence is a good predictor of accuracy (e.g., Winsor et al., in press). Confidence ratings from young children should not necessarily be disregarded as inaccurate. Additionally, *confidence was more predictive of accuracy than age*. Contrary to previous beliefs in eyewitness literature, children as young as 4 seem able to accurately assign confidence judgements to reflect their accuracy. What's more, these confidence judgements were more predictive of accuracy than a child's age. This is an exciting finding, as it provides more evidence for children's confidence being a good predictor of accuracy and suggests that confidence could be considered over age when judging an individual's memory evidence. The key take home message is that, when considering other measures of accuracy, confidence appear to be the best predictor of accuracy. Although confidence being predictive of accuracy in children had already been detected in the developmental literature (e.g., Hembacher & Ghetti, 2014), these studies have typically used basic list learning or perceptual paradigms. The novel element of using a complex episodic event furthers evidence for children's metacognitive ability, as it appears that they can utilise explicit measures on both simple and more complex memory tasks.

In terms of other implicit metacognitive measures, hedges and fillers were predictive of accuracy across experiments. Although we also found other verbal and body gestures–such as head tilts, head nods, and boosters—predicted accuracy, these differed across the two experiments. As such, stable conclusions cannot be made about the usefulness of these in predicating accuracy in children. These measures could be furthered explored in future research with larger sample sizes and other test conditions (such as cued-recall or free recall conditions) before concluding that these are stable predictors of accuracy in children. We must also consider that these measures may be more easily observed in person rather than online.

The evidence presented in this thesis could support emerging evidence that children's memory statements should not be regarded as unreliable. To ascertain the accuracy of children's memory evidence, legal decision makers may be able to implement procedures to assess the relevant metacognitive measures. When further research is conducted, the metacognitive measures found to be predictive of accuracy may be useful to use when ascertaining a child's accuracy in a forensic setting. For example, legal professionals could collect confidence ratings after memory statements from children in this age range and over to help ascertain the reliability of information in their testimonies or to help them determine which information might prove to be the most promising for further investigation. For example, if a child is sure that an unknown perpetrator closed the bedroom curtains, but the child is unsure if the perpetrator drank out of a glass, our research suggests that police resources would be better spent following up with forensic testing on the curtains than on the glass. The box sorting task could also potentially be mimicked in a criminal justice setting: child witnesses could be asked if they want the information they have given kept in a closed-eye box or an open-eye box, with the knowledge that only the information in the open-eye box being considered later.

In closing, children's memory accounts may be underappreciated in the CJS (e.g., Brigham & WolfsKeil, 1983; Goodman, 1984). The result of this scepticism can be devastating, with cases ending in wrongful prosecutions or criminals walking free. This thesis provides support to further emerging evidence that children's memory statements should not be automatically regarded as unreliable, and that legal decision makers could implement procedures using metacognitive measures to help ascertain the accuracy of children's memory evidence. Our results suggest that confidence, response time, box sorting, hedges, and fillers are useful measures when collecting information about a child's accuracy, with confidence providing the most information about accuracy. This work could provide the foundations for bridging the gap between controlled lab studies and real-life legal settings and facilitate further research on child eyewitness metacognition with the potential to have important real-world implications. For example, if these results are replicated in further studies using methods more relevant and applicable to legal settings (such as free and cued recall), legal decision makers could use this information to improve the collection and interpretation of children's memory evidence. A better understanding of children's memory evidence through harnessing metacognition would potentially aid in decreasing the rate of miscarriages of justice and ensure the integrity of the CJS.

References

Ackerman, R., & Koriat, A. (2011). Response latency as a predictor of the accuracy of children’s reports. *Journal of Experimental Psychology: Applied, 4*, 406-417.

Allwood, C. M., Granhag, P. A., & Jonsson, A. C. (2006). Child witnesses' metamemory realism. *Scandinavian Journal of Psychology, 47*(6), 461–470.

Allwood, C. M., Granhag, P. A., & Jonsson, A. C. (2006). Child witnesses’ metamemory realism. *Scandinavian Journal of Psychology, 47*, 461-470.

Almerigogna, J., Ost, J., Bull, R., & Akehurst, L. (2007). A state of high anxiety: How non-supportive interviewers can increase the suggestibility of child witnesses. *Applied Cognitive Psychology, 21*, 963-974.

Armstrong, M. (2020). Children’s epistemic inferences through modal verbs and prosody. *Journal of Child Language*, 1-38.

Asendorpf, J. B., & Baudonnière, P. (1993). Self-Awareness and other- awareness: Mirror self-recognition and synchronic imitation among unfamiliar peers. *Developmental Psychology, 29*, 88-95.

Asendorpf, J. B., Conner, M., Fruyt, F., Houwer, J., Denissen, J. J., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108-119.

Bücker, J., Kapczinski, F., Post, R., Cerésr, K. M., Szobot, C., Yatham, L. N., . . . Kauer-Sant'Anna, M. (2012). Cognitive impairment in school-aged children with early trauma. *Comprehensive Psychiatry, 53*(6), 758-64.

Balcomb, F. K., & Gerken, L. (2008). Three-year old children can access their own memory to guide responses on a visual matching task. *Developmental Sciences, 11*, 750-760.

Bauer, P. J., & Fivush, R. (2014). *The Wiley handbook on the development of children's memory.* Wiley.

Beck, S. R., & Robinson, E. J. (2001). Children’s ability to make tentative interpretations of ambiguous messages. *Journal of Experimental Child Psychology, 79*, 95-114.

Beck, S. R., Robinson, E. J., & Freeth, M. M. (2008). Can children resist making interpretations when uncertain? *Journal of Experimental Child Psychology, 99*, 252-270.

Bell, G. (1986, March 22). 'Child sexual abuse cases doubled last year, NSPCC survey shows.'. *The Times*.

Belletier, C., & Camos, V. (2018). Does the experimenter presence affect working memory?: presence and working memory. *Annals of the New York Academy of Sciences, 00*(1-9), 2-9.

Benton, T. R., Ross, D. F., Bradshaw, E., Thomas, W. N., & Bradshaw, G. S. (2006). Eyewitness memory is still not common sense: Comparing jurors, judges and law enforcement to eyewitness experts. *Applied Cognitive PsychologY, 20*, 115–129.

Beran, M. J., Brandl, J. L., Perner, J., & Proust, J. (2013). On the nature, evolution, development, and epistemology of metacognition: Introductory thoughts. In *Foundations of Metacognition* (pp. 1-26).

Berch, D. B., & Evans, R. C. (1973). Decision processes in children's recognition memory. *Journal of Experimental Child Psychology, 16*(1), 148–164.

Binet, A. (1990). *La suggestibilité.* Paris: Schleicher Feres.

Borràs-Comes, J., Roseano, P., Bosch, M. V., Chen, A., & Prieto, P. (2011). Perceiving uncertainty: facial gestures, intonation, and lexical choice. *Conference on Gesture and Speech in Interaction.*

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). Correlations of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology, 72*, 691-695.

Bowcott, O. (2017, October). Two-year-old girl gives evidence in UK abuse case. *Retrieved from https:// www.theguardian.com/law/2017/oct/10/two-year-old-girl-gives-evidence-in-uk-abuse-case*.

Bowlby, J. (1982). Attachment and loss: Retrospect and prospect. *American Journal of Orthopsychiatry, 52*(4), 664–678.

Brainerd, C. J., & Reyna, V. F. (2012). Reliability of children's testimony in the era of developmental reversals. *Developmental Review, 32*(3), 224–267.

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior, 26*(3), 353–364.

Brewer, N., & Day, K. (2005). The Confidence-Accuracy and Decision Latency-Accuracy Relationships in Children's Eyewitness Identification. *Psychiatry, Psychology and Law, 12*(1), 119-128.

Brewer, N., & Wells, G. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*(12), 11–30. .

Brigham, J. C. (1990). Target person distinctiveness and attractiveness as moderator variables in the confidence-accuracy relationship in facial identification. *Basic and Applied Social Psycholog, 11*, 101- 11.

Brigham, J. C., & WolfsKeil, M. P. (1983). Opinions of attorneys and law enforcement personnel on the accuracy of eyewitness identifications. *Law and Human Behavior*, 337–349.

Brinck, I., & Liljenfors, R. (2013). The developmental origin of metacognition. *Infant and Child Development, 22*, 85-101.

Brown, A. (1983). Learning, remembering and understand- ing. In P. Mussen (Ed.), *Handbook of Child Psychology.* New York: Wiley.

Brown, A. (1987). Metacognition, executive control, self-regulation and other more mysterious systems. In F. E. Weinert, & R. H. Kluwe (Eds.), *Meta- cognition, Motivation and Understanding of Mechanism* (pp. 65-116). Hillsdale, NJ: Erlbaum.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In *Handbook of Child Psychology* (pp. 77-166). New York: Wiley.

Brown, D. A., Lamb, M. E., Lewis, C., Pipe, M., orbach, Y., & Wolfman, M. (2013). The NICHD investigative interview protocol: An analogue study. *Journal of Experimental Psychology: Applied., 19*, 367–382.

Brown, S., & Walker, M. (1983). *The Art of problem Posing.* New York: Franklin Institute Press.

Bruer, K. C., Fitzgerald, R., Price, H. L., & Sauer, J. D. (2017). How sure are you that this is the man you saw? Child witnesses can use confidence judgments to identify a target. *Law and Human Behavio, 41*(6), 1-15.

Bruner, J. S. (1978). The role of dialogue in language acquisition. In A. Sinclair, R. J. Jarvelle, & W. J. Levelt (Eds.), *The Child's Concept of Language.* New York: Springer-Verlag.

Bryce, D., & Whitebread, D. (2012). The development of metacognitive skills: evidence from observational analysis of young children's behavior during problem-solving. *Metacognition Learning, 7*, 197-217.

Bull, R. (2011). The investigative interviewing of children and other vulnerable witnesses: Psychological research and working/professional practice. *Legal and Criminological Psychology*, 5-23.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*.

Calero, M. D., Garcia-Martin, M. B., Jiménez, M. I., Kazén, M., & Araque, A. (2007). Self-regulation advantage for high-IQ children: Findings from a research study. *Learning and Individual Differences, 17*(4), 328-343.

Call, J., & Carpenter, M. (2001). Do apes and children know what they have seen? *Animal Cognition, 4*, 207-220.

Campos, J. J., & Steinberg, C. (1981). Perception, appraisal and emotion: The onset of social referencing. In M. E. Lamb, & L. R. Sherrof (Eds.), *Infant social cognition. Empirical and theoretical considerations* (pp. 273–314). Hillsdale, NJ: Erlbaum.

Campos, J. J., & Stenberg, C. R. (1981). Perception, appraisal, and emotion: The onset of social referencing. In M. E. Lamb, & L. R. Sherrod, *Infant social cognition: Empirical andtheoretical considerations* (pp. 273-314). Hillsdale, NJ: Lawrence Erlbaum Associates.

Carter, H.;. (2012, May). *Victim of Rochdale child sex ring: 'they ripped away all my dignity'.* Retrieved from The Guardian: https://www.theguardian.com/uk/2012/may/08/victim-rochdale-child-sex-ring

Ceci, S. J., & Bruck, M. (1995). Jeopardy in the courtroom: A scientific analysis of children's testimony. *American Psychological Association*.

Ceci, S. J., Ross, D. F., & Toglia, M. P. (1987). Suggestibility of children's memory: Psycholegal implications. *Journal of Experimental Psychology: General, 116*(1), 38-49.

Ceci, S. J., Ross, D. J., & Toglia, M. P. (1989). *Perspectives on children's testimony.* Springer.

Ceci, S., Hritz, A., & Royer, C. (2016). *Understanding suggestibility.*

Chambers, C. T., & Johnston, C. (2002). Developmental differences in children's use of rating scales. *Journal of Pediatric Psychology, 27*(1), 27–36.

Clark, H. H., & Tree, J. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, 73-111.

Collins, K., Harker, N., & Antonopoulos, G. A. (2017). The impact of the registered intermediary on adults' perceptions of child witnesses: Evidence from a mock cross examination. *European Journal on Criminal Policy and Research, 23*, pages 211–225.

Conner, O. L., Siegle, G. J., McFarland, A. M., Silk, J. S., Ladouceur, C. D., Dahl, R. E., . . . Ryan, N. D. (2012). Mom—it helps when you're right here! Attenuation of neural stress markers in anxious youths whose caregivers are present during fMRI. *PLOS One*.

Cooper, P., & Mattison, M. (2017). Itermediaries, vulnerable people and the quality of evidence: An international comparison of three versions of the English intermediary model. *The International Journal of Evidence & Proof.*

Cottrell, N. B., Wack, D., Sekerak, G. J., & Rittle, R. H. (1968). Cottrell et al., 1968. *Journal of Personality and Social Psychology, 9*(3), 245-250.

Crescenzi, A. (2016). Metacognitive knowledge and metacognitive regulation in time-constrained in information search. SAL@ SIGIR.

Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987b). Improving the reliability of eyewitness identifications: Putting context into context. *Journal of Applied Psychology, 72*, 629–637.

Darnell, C. A. (2015). *Children's understanding of uncertainty.* [Unpublished doctoral dissertation]. University of Birmingham.

David, M., & Appell, G. (1961). A study of nursing care and nurse-infant interaction: I. In B. M. Foss, *Dererminants of infant behavior.* London: Methuen.

Davies, E., Hanna, K., Henderson, E., & Hand, L. (2011). *Questioning child witnesses: Exploring the benefits and risks of intermediary models.*

Debras, C. (2017). The shrug: Forms and meanings of a compound enactment. *Gesture, 16*(1), 1-34.

Destan, N., Hembacher, E., Ghetti, S., & Roebers, C. M. (2014). Early metacognitive abilities: the interplay of monitoring and control processes in 5- to 7-year-old children. *Journal of Experimental Child Psychology*.

Dodson, C. S., & Dobolyi, D. G. (2015a). The cross-race effect, decision time and accuracy. *Applied Cognitive Psychology, 30*(1), 113-125.

Donaldson, M. (1978). *Children's Minds.* London: Fontana.

Draeger, S., Prior, M., & Sanson, A. (1986). Visual and auditory attention performance in hyperactive children: Competence or compliance. *Journal of Abnormal Child Psychology volume, 14*, 411–424.

Esken, F. (2012). Early forms of metacognition in human children. In M. J. Beran, J. L. Brandl, J. Perner, & J. Proust (Eds.), *Foundations of Metacognition* (pp. 134-145)). Oxofrd, England: Oxford University Press.

Esposito, A., Marinaro, M., & Palombo, G. (2004). Children speech pauses as markers of different discourse structures and utterance information content. In *From Sound to Sense* (pp. 139 – 144). Cambridge, MA.

Everson, H. T., Smodlaka, I., & Tobias, S. (1994). Exploring the relationship of test anxiety and metacognition on reading test performance: A cognitive analysis. *Anxiety, Stress, and Coping, 7*, 85–96.

Featherstone, L., & Kaladelfos, A. (2016). *Sex crimes in the fifties.*

Featherstone, L., & Kaladelfos, A. (2016). *Sex Crimes in the Fifties.*

Feinman, S. (1992). *Social referencing and the social construction of reality in infancy.* New York: Plenum.

Feinstein, A. R., & Cicchetti, D. V. (1989). High agreement but low kappa: The problems of two paradoxes. *Journal of Clinical Epidemiology, 6*, 543-549.

Finn, B., & Metcalfe, J. (2014). Overconfidence in children's multi-trial judgments of learning. *Learning and Instruction, 32*, 1-9.

Fisher, R. (1998). Thinking about thinking: Developing metacognition in children. *Early Child Development and Care, 141*, 1-15.

Fisher, R., & Geiselman, R. E. (2010). The Cognitive Interview method of conducting police interviews: Eliciting extensive information and promoting therapeutic jurisprudence. *International Journal of Law and Psychiatry, 33*(5-6), 321-328.

Flavell, J. H. (1976). Metacognitive aspects of problem solving. In L. B. (Ed.), & L. B. Resnick (Ed.), *The nature of intelligence* (pp. pp.231-236). Hillsdale, NJ: Erlbaum.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*, 906-911.

Flavell, J. H., & Wellman, H. M. (1977). Perspectives on the development of memory and cognition. In R. V. Kail, & J. W. Hagen (Eds.), *Metamemory.* Hillsdale, NJ: Erlbaum.

Flavell, J. H., Beach, D. R., & Chinsky, J. M. (1966). Spontaneous verbal rehearsal in a memory task as a function of age. *Child Development, 37*, 283-299.

Flavell, J. H., Green, F. L., & Flavell, E. R. (1995). The development of children's knowledge about attentional focus. *Developmental Pscyhology, 31*(4), 706-712.

Flavell, J. H., Green, F. L., & Flavell, E. R. (1995). Young children's knowledge about thinking. *Monographs of the Society for Research in Child Developmen*.

Flavell, J. H., Green, F., & Flavell, E. R. (2000). Development of children's awareness of their own thoughts. *Journal of Cognition and Development, 1*, 97-112.

Flavell, J. H., Miller, P. H., & Miller, S. A. (2002). *Cognitive Development.*

Flavell, J. H., Speer, J. R., Green, F. L., August, D. L., & Whitehurst, G. J. (1981). The development of comprehension monitoring and knowledge about communication. *Monographs of the Society for Research in Child Development, 5*, 1-65.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, 443.

Fox, J. R., Park, B., & Lang, A. (2007). When available resources become negative resources: The effects of cognitive overload on memory sensitivity and criterion bias. *Communication Research, 34*(3), 277-296.

Fritz, K., Howie, P., & Kleitman, S. (2010). "How do I remember when I got my dog?" The structure and development of children's metamemory. *Metacognition Learning, 5*, 207-228.

Fusaro, M., & Harris, P. L. (2013). Dax gets the nod: Toddlers detect and use social cues to evaluate testimony. *Developmental Psychology*.

Fusaro, M., Harris, P. L., & Pan, B. A. (2011). Head nodding and head shaking gestures in children's early communication. *First Language, 32*(4), 439-458.

Fusaro, M., Harris, P. L., & Pan, B. A. (2011). Head nodding and head shaking gestures in children's early communication. *First Language, 32*(4), 439–458.

Fusaro, M., Vallotton, C. D., & Harris, P. L. (2014). Beside the point: Mothers' head nodding and shaking gestures during parent–child play. *Infant Behavior and Development, 37*, 235–247.

Gelman, S. A., & Bloom, P. (2000). Young children are sensitive to how an object was created when deciding what to name it. *Cognition, 76*, 91-103.

Geurten, M., & Bastin, C. (2018). Behaviors speak louder than explicit reports: Implicit metacognition in 2.5-year-old children. *Developmental Science*.

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*, 160-165.

Gill, M. J., Swann, W. B., & Silvera, D. H. (1998). On the genesis of confidence. *Journal of Personality and Social Psychology, 75*, 1101-1114.

Givens, D. (1977). Shoulder shrugging: A densely communicative behavior. *Semiotica, 19*(2), 13-29.

Goldin-Meadow, S. (2015). Gesture and cognitive development. In *Handbook of Child Psychology and Developmental Science* (Vol. 2).

Goodenough, B., Kampel, L., Champion, G. D., Lau-breaux, L., Nicholas, M. K., Ziegler, J. K., & McInerney, M. (1997). n investigation of the placebo effect and age-related factors in the report of needle pain from venepuncture in children. *Pain, 72*, 383-391.

Goodman, G. S. (1984). Children's testimony in historical perspective. *Journal of Social Issues, 40*(2), 9-31.

Goodman, G. S., & Michelli, J. (1981). Would you believe a child witness? *Psychology Today*, 82-95.

Gopnik, A., & Graf, P. (1988). Knowing how you know: Young children's ability to identify and remember the sources of their beliefs. *Child Development,, 59*(5), 1366-1371.

Gould, F., Clarke, J., Heim, C., Harvey, P. D., Majer, M., & Nemeroff, C. B. (2012). The effects of child abuse and neglect on cognitive functioning in adulthood. *Journal of Psychiatric Research, 46*(4), 500–506.

Goupil, L., Romand-Monnier, M., & Kouider, S. (2016). Infants ask for help when they know they don’t know. *PNAS, 113*(13), 3492-3496.

Guerin, B. (1986). Mere presence effects in humans: a review. *Journal of Experimental Social Psychology, 22*, 38-77.

Gustafsson, P. U., Lindholm, T., & Jönsson, F. U. (2019). Predicting accuracy in eyewitness testimonies with memory retrieval effort and confidence. *Frontiers in Psychology, 10*, 1-10.

Gustafsson, P. U., Lindholm, T., & Jönsson, F. U. (2019). Predicting accuracy in eyewitness testimonies with memory retrieval effort and confidence. *Frontiers*.

Gwet, K. L. (2001). *Handbook of Inter-Rater Reliability: How to Estimate the Level of Agreement Between Two or Multiple Raters.* Gaithersburg, MD: STATAXIS Publishing Company.

Gwet, K. L. (2002). Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. In *Statistical Methods For Inter-Rater Reliability Assessmen.*

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*, 29–48.

Gwet, K. L. (2014). *Benchmarking Agreement Coefficients*. Retrieved from K. Gwet's Inter-Rater Reliability Blog: https://inter-rater-reliability.blogspot.com/2014/12/benchmarking-agreement-coefficients.html

Hübscher, I., Esteve-Gilbert, N., & Igualada, A. (2017). Intonation and gesture as bootstrapping devices in speaker uncertainty. *First Languag, 37*(1), 24-41.

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*(4).

Händel, M., & Fritzsche, E. S. (2014). Students’ confidence in their performance judgements: a comparison of different response scales. *Educational Psychology: An International Journal of Experimental Educational Psychology, 35*(3), 377-395.

Hamilton, M. H., & Addison, G. (1947). *Criminal law and procedure: Containing the Crimes Ac.* Sydney, NSW, Australia.

Hanway, P., Akehurst, L., Vernham, Z., & Hope, L. (2020). The effects of cognitive load during an investigative interviewing task on mock interviewers’ recall of information. *Legal and Criminological Psychology, 26*(1), 25-41.

Harris, P. L. (2005). Conversation, pretense and theory of mind. In J. W. Astington, & J. A. Baird (Eds.), *Why language matters for theory of mind.* (pp. 70–83). New York: Oxford University Press.

Harris, P. L., Bartz, D. T., & Rowe, M. L. (2017). Young children communicate their ignorance and ask questions. *PNAS, 114*(30), 7884-7891.

Harris, P. L., Bartz, D. T., & Rowe, M. L. (2017). Young children communicate their ignorance and ask questions. *PNAS, 114*(38).

Harry-Augstein, S., & Thomas, L. (1991). *Learning Conversation.* London: Routledge.

Hembacher, E., & Ghetti, S. (2014). Don't look at my answer: Subjective uncertainty underlies preschoolers' exclusion of their least accurate memories. *Psychological Science, 25*(9), 1768–1776.

Henry, L. A., Crane, L., Nash, G., Hobson, Z., Kirke-Smith, M., & Wilcock, R. (2017). Verbal, Visual, and Intermediary Support for child witnesses with autism during investigative interviews. *Journal of Autism and Developmental Disorders volume, 47*, 2348–2362.

Hiller, R. M., & Weber, N. (2013). A comparison of adults' and children's metacognition for yes/no recognition decisions. *Journal of Applied Research in Memory and Cognition, 2*(3), 185–191.

Hofer, B. K. (2004). Epistemological understanding as a metacognitive process: Thinking aloud during online searching. *Educational Psychologist, 39*(1), 43-55.

Holliday, R. E., Brainerd, C. J., Reyna, V. F., & Humphries, J. E. (2009). The Cognitive Interview. In R. Bull, & T. Williamson (Eds.), *andbook of the psychology of investigative interviewing* (pp. 137-160). Chichester: Wiley-Blackwell.

Holmes, J. (1990). Hedges and boosters in women's and men's speech. *Language % Communicaoon., 10*(3), 185-205.

Home Affairs Committee. (2013). *Child sexual exploitation and the response to localised grooming.* House of Commons, Home Affairs Committee, London.

Horry, R., Halford, P., Brewer, N., Milne, R., & Bull, R. (2014). Archival analyses of eyewitness identification test outcomes: What can they tell us about eyewitness memory? *Law and Human Behavior, 1*, 94-108.

Humphries, J. E., Holliday, R. E., & Flowe, H. D. (2012). Faces in motion: Age-related changes in eyewitness identification performance in simultaneous, sequential, and elimination video lineups. *Applied Cognitive Psychology, 26*, 149–158.

Hyland, K. (1998). Boosting, hedging and the negotiation of academic knowledge. *Text & Talk*, 349-382.

Istomina, Z. M. (1975). The development of volutary memory in preschool-age children. *Soviet Psychology, 13*, 5-64.

Jaffee, S. R., & Maikovich-Fong, A. K. (2010). Effects of chronic maltreatment and maltreatment timing on children's behavior and cognitive abilities. *Journal of Child Psychology and Psychiatry, 52*(2), 84–194.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *ournal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1304–1316.

Kanfer, R., & Ackerman, P. L. (1996). A self-regulatory skills perspective to reducing cognitive interference. In I. G. Sarason, G. R. Pierce, & B. R. Sarason (Eds.), *Cognitive interference: Theories, methods, and findings* (pp. 153-171). Erlbaum.

Kassin, S. M., & Kiechel, K. L. (1996). The social psychology of false confessions: Compliance, internalization, and confabulation. *Psychological Science*.

Katz, S., & Mazur, A. M. (1979). *Understanding the rape victim.* New York: Wiley.

Keast, A., Brewer, N., & Wells, G. L. (2007). Children's metacognitive judgments in an eyewitness identification task. *Journal of Experimental Child Psychology, 97*, 286–314.

Kebbell, M. R., & Johnson, S. D. (2000). Lawyers' questioning: The effect of confusing questions on witness confidence and accuracy. *Law and Human Behavior volume, 24*, 629–641.

Keeney, T. J., Cannizzo, S. R., & Flavell, J. H. (1967). Spontaneous and induced verbal rehearsal in a recall task. *Child Development, 38*(4), 953-966.

Keeney, T. J., Cannizzo, S. R., & Flavell, J. H. (1967). Spontaneous verbal rehearsal in a memory task as a function of age. *Child Develop, 38*(4), 953-966.

Kendon, A. (2003). Some uses of the head shake. *Gesture, 2*(2), 147–182.

Kendon, A. (2004). *Gesture: Visible action as utterance.* Cambridge University Press .

Kim, G., & Kwak, K. (2011). Uncertainty matters: impact of stimulus ambiguity on infant social referencing. *Infant Child Development, 20*, 449–463.

Kim, S., Paulus, M., Sodian, B., & Proust, J. (2016). Young children's sensitivity to their own ignorance in informing others. *PLoS ONE, 11*(3).

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences, 17*(2), 161-173.

Kloo, D., & Rohwer, M. (2012). The development of earlier and later forms of metacognitive abilities: reflections on agency and ignorance. In M. J. Beran, J. L. Brandl, J. Perner, & J. Proust (Eds.), *Foundations of Metacognition* (pp. 167-180). Oxford: Oxford University Press.

Knutsson, J., & Allwood, C. M. (2014). Opinions of legal professionals: Comparing child and adult witnesses' memory report capabilities. *The European Journal of Psychology Applied to Legal Context, 6*, 79-89.

Koriat, A. (2000). The feeling of knowing: Some Meta-theoretical implications for consciousness and control. *Consciousness and Cognition, 9*, 149-171.

Koriat, A., & Ackerman, R. (2010). Choice latency as a cue for children's subjective confidence in the correctness of their answers. *Developmental Science, 13*(3), 441-453.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*(3), 490-517.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*(3), 490-517.

Koriat, A., & Levy-Sadot, R. (1999). Processes underlying metacognitive judgements: Information-based and experienced based monitoring of ones' own knowledge. In S. Chaiken, & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 483-502). Guilford: New York Publications.

Koriat, A., Goldsmith, M., Scheider, W., & Nakash-Dura, M. (2001). The credibility of children's testimony: Can children control the accuracy of their memory reports? *Journal of Experimental Child Psychology, 79*, 405–437.

Krahmer, E., & Swerts, M. (2005). How children and adults produce and perceive uncertainty in audiovisual speech. *Language and Speech, 48*(1), 29-53.

Krebs, S. S., & Roebers, C. M. (2010). Children's strategic regulation, metacognitive monitoring, and control processes during test taking. *British Journal of Educational Psychology, 80*, 325–340.

Kuhn, D. (1999). A Developmental Model of Critical Thinking. *Educational Researcher, 28*(26).

Kurdi, B., Diaz, A. J., Wilmuth, C. A., Friedman,, M. C., & Banaji, M. R. (2017). Variations in the relationship between memory confidence and memory accuracy: The effects of spontaneous accessibility, list length, modality, and complexity. *Psychology of Consciousness: Theory, Research, and Practice, 17*, 2326-5523.

Lakoff, R. (1975). Linguistic theory and the real world. *Language Learning, 25*(2), 309-338.

Leckey, S., Selmeczy, D., Kazemi, A., Johnson, E. G., Hembacher, E., & Ghetti, S. (2020). Response latencies and eye gaze provide insight on how toddlers gather evidence under uncertainty. *Nature Human Behaviour, 4*, 928–936.

Leippe, M. R., Manion, A. P., & Romanczyk, A. (1992). Eyewitness persuasion: How and how well do fact finders judge the accuracy of adults' and children's memory reports? *Journal of Personality and Social Psychology, 63*(2), 181-197.

Lipko, A. R., Dunlosky, J., & Merriman, W. E. (2009). Persistent overconfidence despite practice: The role of task experiences in pre-schoolers' recall predictions. *Journal of Experimental Child Psychology, 103*, 152-166.

Lipko, A. R., Dunlosky, J., Lipowski, S. L., & Merriman, W. E. (2012). Young children are not under-confident with practice: The benefit of ignoring a 211 References fallible memory heuristic. *Journal of Cognition and Development, 13*, 174-188.

Lockl, K., & Scheider, W. (2002). Developmental trends in children's feeling-of-knowing judgements. *International Journal of Behavioral Development.*

Loftus, E. F. (1994). The repressed memory controversy. *American Psychologist, 49*(5), 443–445.

Lohman , H., & Tomasello, M. (2003). he role of language in the development of false belief understanding: a training study. *Child Development, 74*, 1130–1144.

Lyons, K. E., & Ghetti, S. (2011). The Development of uncertainty monitoring in early childhood. *Child Development, 82*(6), 1778–1787.

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.

MacWhinney, B., & Osser, H. (1977). Verbal planning functions in children's speech. *Child Development, 48*(3), 978–985.

Mahmoud, M., & Robinson, P. (2011). Interpreting hand-over-face gestures. *International Conference on Affective Computing and Intelligent Interaction*, (pp. 248-255).

Mansour, J. K., Beaudry, J. L., & Lindsay, R. C. (2017). Are multiple-trial experiments appropriate for eyewitness identification studies? Accuracy, choosing, and confidence across trials. *Behavioural Research*.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*, 537-563.

McGough, L. S. (1993). *Child witnesses: Fragile voices in the American legal system.* New Haven, CT: Yale University Press.

Melinder, A., Goodman, G. S., Eilertsen, D. E., & Magnussen, S. (2004). Beliefs about child witnesses: A survey of professionals. *Psychology, Crime & Law, 10*(4), 347-365.

Memon, A., Meissner, C. A., & Fraser, J. (2010). The Cognitive Interview: A meta-analytic review and study space analysis of the past 25 years. *Psychology, Public Policy, and Law*, 340–372.

Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. (n.d.). *American Psychologist*, 34, 906 - 911.

Metcalfe, J., & Finn, B. (2013). Metacognition and control of study choice in children. *Metacognition Learning, 8*, 19–46.

Metcalfe, J., & Shimamura, A. P. (1994/1996). *Metacognition. Knowing about knowing.* Cambridge, MA: MIT Press.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition, 4*, 93–102.

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General,, 140*, 239–257.

Mickes, L., Moreland, M. B., Clark, S. E., & Wixted, J. T. (2014). Missing the information needed to perform ROC analysis? Then compute d′, not the diagnosticity ratio. *Journal of Applied Research in Memory and Cognition, 3*(2), 58–62.

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review.*

Ministry of Justice. (2011, March). Achieving best evidence in criminal proceedings: Guidance on interviewing victims and witnesses, and guidance on using special measures. Retrieved from https://www.cps.gov.uk/sites/default/files/documents/legal_guidance/best_evidence_in_criminal_proceedings.pdf

Monosov, I. E., & Hikosaka, O. (2013). Selective and graded coding of reward uncertainty by neurons in the primate anterodorsal septal region. *Nature neuroscience, 16*, 756–762.

Moore, C., Bryant, D., & Furrow, D. (1989). Mental terms and the development of certainty. *Child Development, 60*(1), 167-171.

Neil v. Biggers, 409 (U.S. 188 1972).

Nelson, T. O., & Dunlosky, J. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5*(4), 207–213.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In *The Psychology of Learning and Motivation* (Vol. 26). Academic Press. Inc.

Nemeroff, C. B. (2004). Neurobiological consequences of childhood trauma. *The Journal of Clinical Psychiatry, 65*(1), 18-28.

Newcombe, P. A., & Bransgrove, J. (2007). Perceptions of witness credibility: Variations across age. *Journal of Applied Developmental Psychology, 28*, 318-331.

Nguyen, C. D., Carlin, J. B., & Lee, K. (2017). Model checking in multiple imputation: an overview and case study. *Emerging Themes in Epidemiology*.

Nilsen, E. S., Graham, S. A., & Chambers, C. G. (2008). Pre-schooler's sensitivity to referential ambiguity: evidence for dissociation between implicit understanding and explicit behaviour. *Developmental Science, 11*, 556-562.

Office for National Statistics. (2019). *Child abuse and the criminal justice system, England and Wales: year ending March 2019*. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/childabuseandthecriminaljusticesystemenglandandwales/yearendingmarch2019

Office for National Statistics. (2020). *'Characteristics of children in need' National Statistics*. Retrieved from eexplore-education-statistics.service.gov.uk: https://explore-education-statistics.service.gov.uk/find-statistics/characteristics-of-children-in-need/2020

Papaleontiou-Louca, E. (2003). The concept and instruction of metacognition. *Teacher Development, 71.*

Patterson, C. J., Cosgrove, J. M., & O'Brien, R. G. (1980). Nonverbal indicants of comprehension and noncomprehension in children. *Developmental Psychology, 16*(1), 39-48.

Penrod, S., & Cutler, B. (1990). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law, 1*(4), 817–845.

Perner, J. (2012). MiniMeta: in search of minimal criteria for metacognition. In M. J. Beran, J. L. Brandl, J. Perner, & J. Proust, *Foundations of Metacognition* (pp. 234-251). Oxford, England: Oxford University Press.

Perner, J. (2012). MiniMeta: in search of minimal criteria for metacognition. In M. J. Beran, J. L. Brandl, J. Perner, & J. Proust (Eds.), *Foundations of Metacognition* (pp. 234-251). Oxford: Oxford University Press.

Peterson, C., & Briggs, M. (2001). "I was really, really, really mad!" Children's use of evaluative devices in narratives about emotional events. *Sex Roles, 45*(11), 801-825.

Piaget, J. (1950). *The Psychology of Intelligence.* London: Routledge.

Piaget, J. (1964/1968). *Six psychological studies.* New York: Random House.

Pidd, H., & Dodd, V. (2020, Jan). Retrieved from The Guardain: https://www.theguardian.com/uk-news/2020/jan/14/police-errors-may-have-let-abusers-of-up-to-52-children-escape-justice

Pillow, B. H., & Anderson, K. L. (2006). Children's awareness of their own certainty and understanding of deduction and guessing. *British Journal of Developmental Psychology, 24*, 823-849.

Platania, J., & Moran, G. (2010). Social facilitation as a function of the mere presence of others. *The Journal of Social Psychology, 141*(2), 190-197.

Plonikoff, J., & Woolfson, R. (2011). *Young witnesses in criminal proceedings. A progress report on Measuring up?* London: Nuffield Foundation.

Plotnikoff, J., & Woolfson, R. (2007). *The 'Go-Between': evaluation of intermediary pathfinder projects.* Lexicon Limited.

Poulin-Dubois, D., & Brosseau-Liard, P. (2016). The developmental origins of selective social learning. *Current Directions in Psychological Science, 25*, 60–64.

Poulin-Dubois, D., Sodian, B., Metz, U., Tilden, J., & Schoeppner, B. (2007). Out of sight is not out of mind: Developmental changes in infants' understanding of visual perception during the second year. *Journal of Cognition and Development, 8*, 401–21.

Powell, M. B., Garry, M., & Brewer, N. (2013). Expert evidence: Law, practice, procedure and advocacy (5th ed.). In I. Freckelton, & H. Selby (Eds.), *Expert evidence: Law, practice, procedure and advocacy.* Sydney, Australia: Thomson Reuters.

Powell, M. B., Garry, M., & Brewer, N. (2013). Eyewitness testimony. In I. Freckelton, & H. Selby, *Expert evidence: Law, practice, procedure and advocacy* (5th ed.). Sydney, Australia: Thomas Reuters.

Powell, M., Garry, M., & Brewer, N. (2013). Eyewitness testimony. In I. R. Freckelton, & H. Selby (Eds.), *Expert Evidence: Law, Practice, Procedure and Advocacy.* Pyrmont, N.S.W.: Thomson Reuters.

Pratt, C., & Bryant, P. (1990). Young children understand that looking leads to knowing (so long as they are looking into a single barrel). *Child Development, 6*, 973-982.

Pressley, M., Levin, J. R., & Ghatala, E. (1984). Memory strategy monitoring in adults and children. *Journal of Verbal Learning and Verbal Behavior, 23*, 270–288.

Pressley, M., Levin, J. R., Ghatala, E. S., & Ahmad, M. (1987). Test monitoring in young grade school children. *Journal of Experimental Child Psychology,, 43*, 96-111.

Proust, J. (2007). Metacognition and meta-representations: Is self-directed theory of mind a precondition for metacognition? *Synthese,, 2*, 271-295.

Proust, J. (2010). Metacognition. *Philosophy Compass, 5*(11), 989–998.

Proust, J. (2012). Metacognition & mindreading: One or two functions? In M. J. Beran, J. L. Brandl, J. Perner, & J. Proust (Eds.), *Foundations of Metacognition* (pp. 234-251). Oxford: Oxford University Press.

R v Wallwork: CCA 1958, 42 (1958).

Reder, L. M., & Schunn, C. D. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. M. Reder (Ed.), *Implicit Memory and Metacognition.* Mahwah, NJ: Erlbaum.

Reid, V., Hoehl, S., & Striano, T. (2006). The perception of biological motion by infants: An event-related potential study. *Neuroscience Letters*, 211–214.

Robinson, E. J., & Whitaker, S. J. (1985). Children's responses to ambiguous messages and their understanding of ambiguity. *Developmental Psychology, 21, 446-454., 21*, 446-454.

Roderer, T., & Roebers, C. M. (2010). Explicit and implicit confidence judgments and developmental differences in metamemory: an eye-tracking approach. *Metacognition Learning, 5*, 229–250.

Roebers, C. M. (2002). Confidence judgments in children's and adult's event recall and suggestibility. *Developmental Psychology, 38*, 1052-1067.

Roebers, C. M. (2002). onfidence judgments in children's and adult's event recall and suggestibility. *Developmental Psychology, 38*, 1052-1067.

Roebers, C. M. (2013). Children's deliberate memory development: The contribution of strategies and metacognitive processes. In *The Wiley Handbook on the Development of Children's Memory.*

Roebers, C. M., & Howie, P. (2003). Confidence judgements in event recall: Developmental progression in the impact of question format. *Journal of Experimental Child Psychology, 85, 352-371*(85), 352-371.

Roebers, C. M., Kälin, S., & Aeschlimann, E. A. (2019). A comparison of non-verbal and verbal indicators of young children's metacognition. *Metacognition and Learning*.

Roebers, C. M., von der Linden, N., & Howie, P. (2007). Favourable and unfavourable conditions for children's confidence judgements. *British Journal of Developmental Psychology, 25*, 109-134.

Roebers, C., Schmid, C., & Roderer, T. (2009). Metacognitive monitoring and control processes involved in primary school children's test performance. *British Journal of Educational Psychology, 79*(4), 749–767.

Roseano, P., González, M., Borràs-Comes, J., & Prieto, P. (2014). Communicating epistemic stance: How speech and gesture patterns reflect epistemicity and evidentiality. *Discourse Processe*, 1-40.

Ross, D. F., Dunning, D., Tonglia, M. P., & Ceci, S. J. (1990). The child in the eyes of the jury: Assessing mock jurors' perceptions of the child witness. *Law and Human Behaviour, 14*, 5–23.

Rozell, S. (1985). Are Children Competent Witnesses?: A Psychological Perspective. *Washington University Law Review, 63*(4), 815-829.

Sauer, J. D., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*, 337–347.

Scaife, M., & Bruner, J. S. (1975). The capacity for joint visual attention in the infant. *Nature, 253*, 265–266.

Scheider, W., Visé, M., Lockl, K., & Nelson, T. O. (2000). Developmental trends in children's memory monitoring: Evidence from a judgment-of-learning task. *Cognitive Development, 15*, 115-134.

Schmid, J., Herholz, S. C., Brandt, M., & Buchner, A. (2010). Recall-to-reject: The effect of category cues on false recognition. *Memory*, 863-882 .

Schneider, W., & Artelt, C. (2010). Metacognition and mathematics education. *ZDM Mathematics Education, 42*, 149–161.

Schneider, W., & Lockl, K. (2002). The development of metacogni- tive knowledge in children and adolescents. In T. J. Perfect, & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 224–257). New York, NY: Cambridge University Press.

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*(4), 351–371.

Schwarz, N., Knäuper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*, 570-582.

Seale-Carlisle, T. M., Colloff, M. F., Flowe, H. D., Wells, W., Wixted, J. T., & Mickes, L. (2019). Confidence and response time as indicators of eyewitness identification accuracy in the lab and in the real world. *Journal of Applied Research in Memory and Cognition, 8*(4), 420-428.

Shore, B. M., & Dover, A. C. (1987). Metacognition, intelligence and giftedness. *Gifted Child Quarterly, 31*(1), 37-39.

Smith, J. D., Shields, W. E., & Washburn, D. A. (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioural and Brain Sciences, 26*, 317–373.

Smith, L. (1994). The development of modal understanding: Piaget's possibility and necessity. *New Ideas in Psychology, 12*, 75-87.

Smith, L. E., & Crabbe, J. (1976). Experimenter role relative to social facilitation and motor learning. *International Journacl of Sport Psychology, 7*, 158-168.

Smith, M.;. (2013, May). Retrieved from The Guardian: https://www.theguardian.com/uk/2013/may/24/rochdale-failures-child-abuse-gang

Sodian, B., Thoermer, C., Kristen, S., & Perst, H. (2012). Metacognition in infants and young children. In M. J. Beran, J. L. Brandl, J. Perner, & J. Proust (Eds.), *Foundations of Metacognition* (pp. 119–133). Oxford University Press.

Son, L. K., & Schwartz, B. L. (2002). The relation between metacognitive monitoring and control. In T. J. Perfect, & B. S. Schwartz (Eds.), *Applied Metacognition* (pp. 15-38). Cambridge: Cambridge University Press.

Spearing, E. R., & Wade, K. A. (2021). Providing eyewitness confidence judgements during versus after eyewitness interviews does not affect the confidence-accuracy relationship. *Journal of Applied Research in Memory and Cognition*.

Statista Research Department. (2021). *Number of child abuse offences in England and Wales 2002-2021*. Retrieved from https://www.statista.com/statistics/303514/child-cruelty-abuse-in-england-and-wales-uk-y-on-y/

Stern, W. (1910). Abstracts of lectures on the psychology of testimony and the study of individuality. *American Journal of Pscyhology, 21*, 270-282.

Stotland, E., & Zander, A. (1958). Effects of public and private failure on self-evaluation. *The Journal of Abnormal and Social Psychology, 56*(2), 223-229.

Streeck, J. (2009). *Gesturecraft: The manufacture of meaning.* John Benjamins Publishing Compan.

Swanson, H. L., Kehler, P., & Jerman, O. (2010). Working memory, strategy knowl- edge, and strategy instruction in children with reading disabilities. *Journal of Learning Disabilities, 43*.

Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language, 53*, 81-94.

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications, 2*(1), 49.

Tekin, E., Lin, W., & Roediger III, H. L. (2018). The relationship between confidence and accuracy with verbal and verbal + numeric confidence scales. *Cognitive Research: Principles and Implications, 3*(14).

Tekin, E., Wenbo, L., & Roediger III, H. L. (2018). The relationship between confidence and accuracy with verbal and verbal and numeric confidence scales. *Cognitive Research: Principles and Implication, 41*.

Veenman, M. V., Kerseboom , L., & Imthorn, C. (2000). Test anxiety and metacognitive skillfulness: Availability versus production deficiencies. *Anxiety, Stress, & Coping: An International Journal, 13*(4), 391-412.

Veenman, M. V., Van Hout-Wolters, B., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1*(1), 3-14.

Vincze, L., & Poggi, I. (2016). I am definitely certain of this! Towards a multimodal repertoire of signals communicating a high degree of certainty. *Nordic Symposium on Multimodal Communication .*

Visser, M., Krahmer, E., & Swerts, M. (2014). Children's expression of uncertainty in collaborative and competitive contexts. *Language and Speech, 57*(1), 86-107.

von Baeyer, C. L., & Webb, G. C. (1997). Underprediction of pain in children undergoing ear piercing. *Behaviour Research and Therapy, 35*(5), 399-404.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes.* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.) Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1962). Thought and Word. In L. Vygotsky, E. Hanfmann, & G. Vakar (Eds.), *Thought and Language* (pp. 119–153). MIT Press.

Wade, A. E. (1997). The child witness and the criminal justice process : A case study in law reform. *Published doctoral dissertation, University of Leeds*. White Rose eTheses Online.

Wassmer, E., Minnaar, G., Abdel, N., Atkinson, M., Gupta, E., Yuen, S., & Rylance, G. (2004). How do paediatricians communicate with children and parents? *Acta Pæditrica, 93*, 1501-1506.

Waters, G. M. (2009). *The limits of young children"s understanding of sources of knowledge.* [Unpublished doctoral dissertation]. University of Birmingham.

Wellman, H. M. (1977). Tip of the tongue and feeling of knowing experiences: A developmental study of memory monitoring. *Child Development, 48*, 13-21.

Wells, G. L. (2014). *The Houston Police Department Eyewitness Identification Experiment: Analysis and Results*. Retrieved from www.lemitonline.org/research/projects.html.

Wells, G. L., & Murray, D. M. (1983). What can psychology say about the Neil v. Biggers criteria for judging eyewitness accuracy? *Journal of Applied Psychology, 68*(3), 347–362.

Wells, G. L., & Murray, D. M. (1984). Eyewtiness confidence. In G. L. Wells, & E. F. Loftus (Eds.), *Eyewitness Testimony: Psychological Perspective* (pp. 155-170). New Yrok, NY: Cambridge University Press.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22*(6), 603-647.

Weston, H. E., Boxer, B., & Heatherington, L. (1998). Children's attributions about family arguments: Implications for family therapy. *Family Process, 37*, 35-49.

Whipple, G. M. (1909). The observer as reporter: A survey of the psychology of testimony. *Psychological Bulletin, 153–170*, 1909.

Whitebread, D., Coltman, P., Pastermak, D., Sangster, C., Grau, V., & Bingham, C. (2009). he development of two observational tools for assessing metacognition and self-regulated learning in young children. Metacognition and Learning,. *Metacognition and Learning, 4*(1), 63–85.

Wigmore, J. H. (1935/1976). *Evidence in trials at common law (revised by J. Chadborn)* (Vol. 6). Boston, MA: Little. Brown & Co.

Winne, P. H. (1996). A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences, 8*(4), 327–353.

Winsler, A., Fernyhough, C., & Montero, I. (n.d.). *Private speech, executive functioning, and the development of verbal self-regulation.* Cambridge: Cambridge University Press.

Winsor, A., Flowe, H. D., Seale-Carlilse, T., Kileen, I., Hett, D., Jores, T., . . . Colloff, M. F. (in press). Child witness expressions of certainty are informative.

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger III, H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychology, 70*(6), 515-26.

Wixted, J., & Wells, G. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science, 18*(1), 10-65.

Wright, D. B., & London, K. (2010). Multilevel modelling: Beyond the basic applications. *British Journal of Mathmatical and Statistical Psychology, 62*(2).

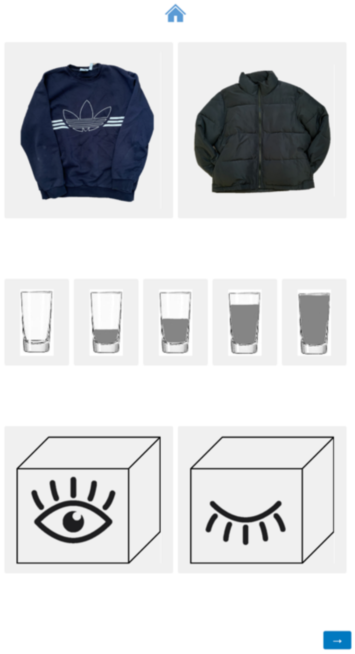
Yarmey, A. D., & Jones, H. P. (1983). Is the psychology of eyewitness identification a matter of common sense'? In S. M. Loyd, & B. R. Clifford (Eds.), *Evuluuting witness evidence* (pp. 13-40). New York: Wiley.

Zajac, R., Westera, N., & Kaladelfos, A. (2008). The "Good Old Days" of Courtroom Questioning: Changes in the Format of Child Cross-Examination Questions Over 60 Years. *Child Maltreatment, 23*(2), 186-195.

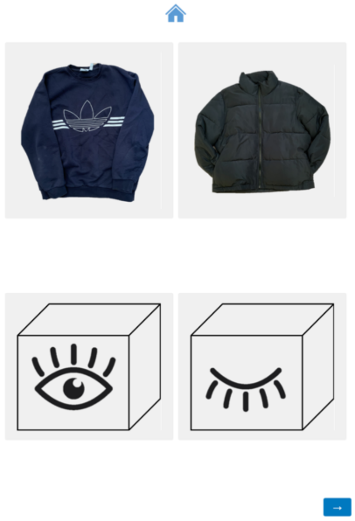
Zeidner, M. (1998). *Test anxiety: The state of the art.* Plenum Press.

Appendices

Appendix A – Example question presentation on Qualtrics for Experiment 1 and Experiment 2 (confidence present condition)



Appendix B – Example question presentation on Qualtrics for Experiment 2 (confidence absent condition)



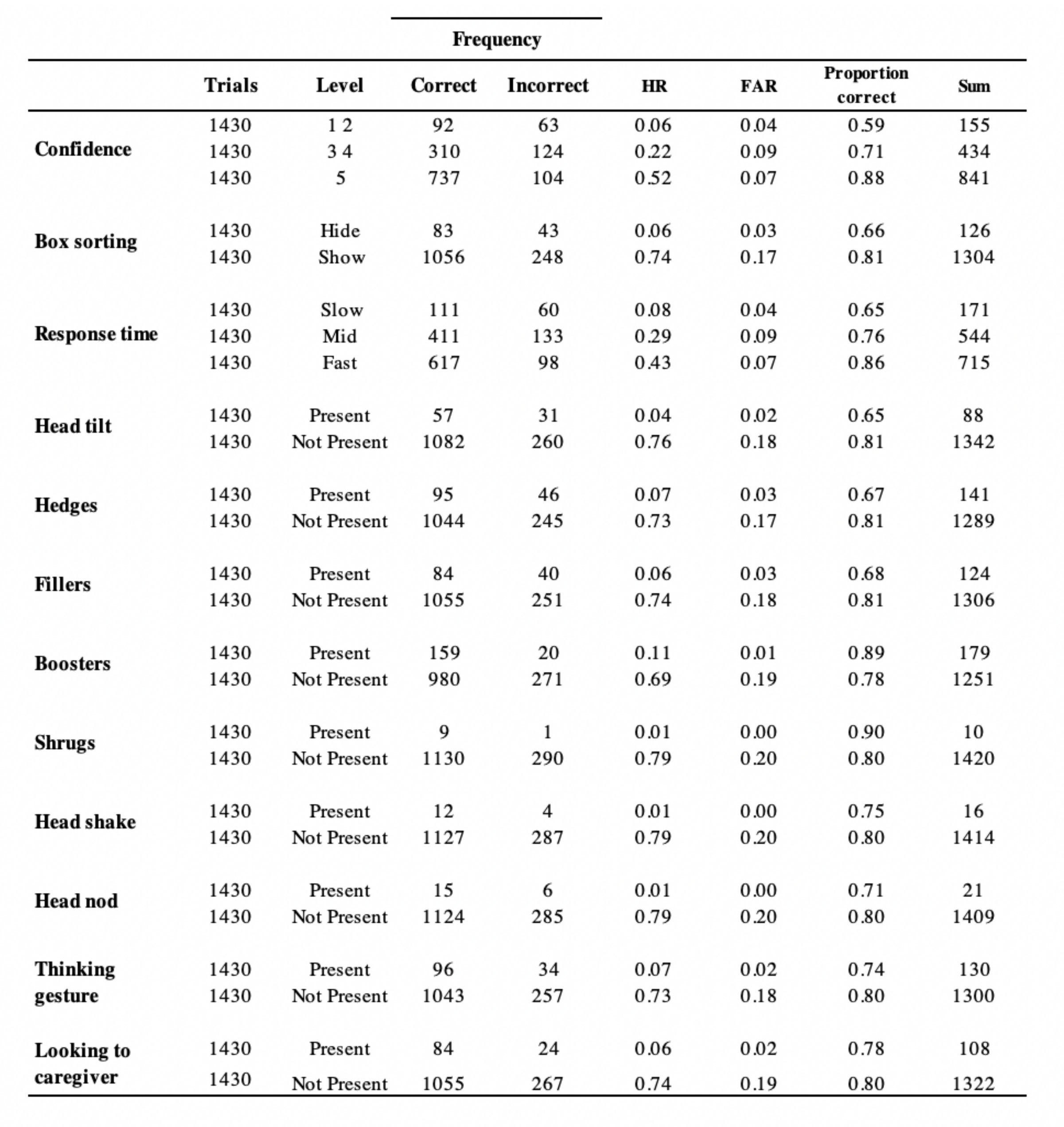
Appendix C – Questions for video 1 (making breakfast)

1. What was on the draining board next to the sink?
2. What was on the windowsill?
3. Which kettle was in the kitchen?
4. Which drink did the boy choose?
5. Which food did the boy choose?
6. What did the fridge look like?
7. What was the boy wearing?
8. Which top was the boy wearing?
9. Which bowl did the boy use?
10. What did the cupboards look like?
11. Which glass did he use?
12. What colour were the counter tops?
13. Which crisps were in the background?
14. What did the microwave look like?
15. What did the boy get out of the fridge?
16. What was on top of the microwave?
17. What did the boy get out of the drawer?
18. What did the sugar pot look like?
19. What did his glasses look like?
20. What else was in the fridge?

Appendix D - Questions for video 2 (washing up)

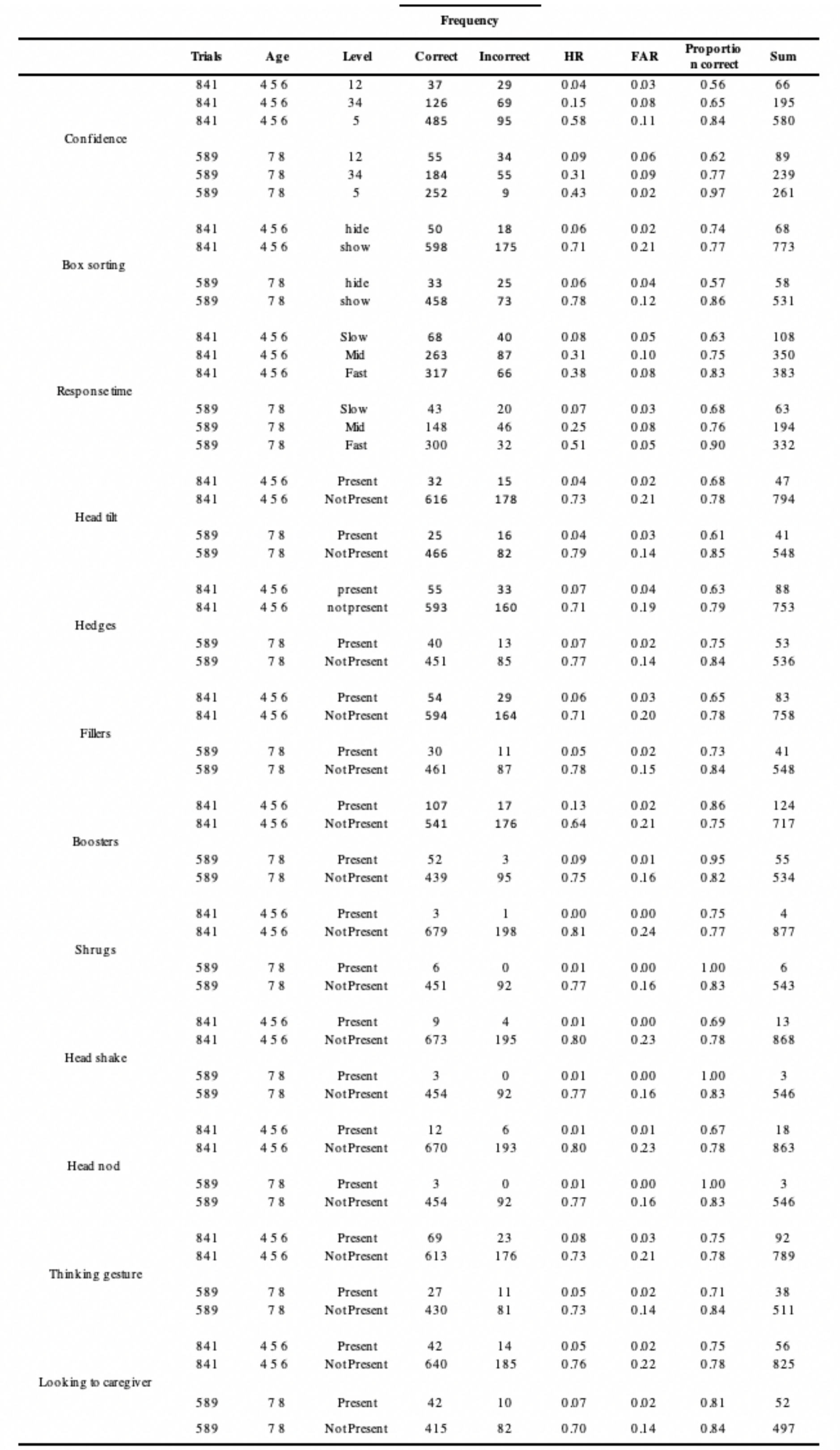
1. What colour was the soap?
2. What was on the windowsill?
3. Which shoes was the boy wearing?
4. What did the bin look like?
5. What did the tea towel look like?
6. What was next to the bin?
7. Which bag was on the countertop?
8. Which cup did he wash up?
9. Which top was he wearing?
10. Which hat was he wearing?
11. Were the blinds open or closed?
12. What did the floor look like?
13. Which vase was in the kitchen?
14. What colour were the walls?
15. What colour was the radiator?
16. What did he use to wash up?
17. What did the washing machine look like?
18. Which socks was he wearing?
19. What was on the wall?
20. What did the tap look like?

Appendix E – Frequency of measures for Experiment 1 across all ages and trials



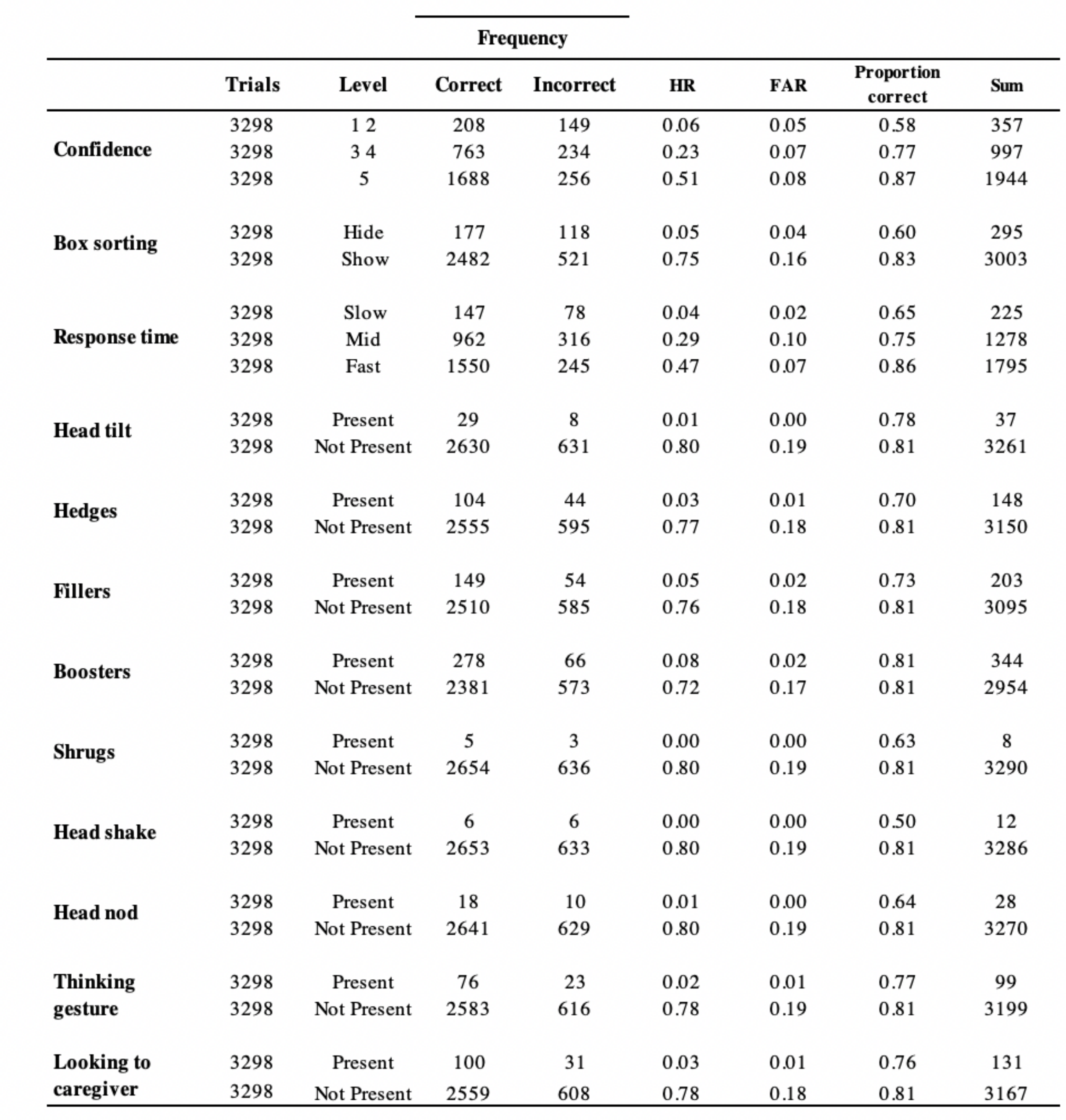
			Frequency					
	Trials	**Level**	**Correct**	**Incorrect**	**HR**	**FAR**	**Proportion correct**	**Sum**
Confidence	1430	1 2	92	63	0.06	0.04	0.59	155
	1430	3 4	310	124	0.22	0.09	0.71	434
	1430	5	737	104	0.52	0.07	0.88	841
Box sorting	1430	Hide	83	43	0.06	0.03	0.66	126
	1430	Show	1056	248	0.74	0.17	0.81	1304
Response time	1430	Slow	111	60	0.08	0.04	0.65	171
	1430	Mid	411	133	0.29	0.09	0.76	544
	1430	Fast	617	98	0.43	0.07	0.86	715
Head tilt	1430	Present	57	31	0.04	0.02	0.65	88
	1430	Not Present	1082	260	0.76	0.18	0.81	1342
Hedges	1430	Present	95	46	0.07	0.03	0.67	141
	1430	Not Present	1044	245	0.73	0.17	0.81	1289
Fillers	1430	Present	84	40	0.06	0.03	0.68	124
	1430	Not Present	1055	251	0.74	0.18	0.81	1306
Boosters	1430	Present	159	20	0.11	0.01	0.89	179
	1430	Not Present	980	271	0.69	0.19	0.78	1251
Shrugs	1430	Present	9	1	0.01	0.00	0.90	10
	1430	Not Present	1130	290	0.79	0.20	0.80	1420
Head shake	1430	Present	12	4	0.01	0.00	0.75	16
	1430	Not Present	1127	287	0.79	0.20	0.80	1414
Head nod	1430	Present	15	6	0.01	0.00	0.71	21
	1430	Not Present	1124	285	0.79	0.20	0.80	1409
Thinking gesture	1430	Present	96	34	0.07	0.02	0.74	130
	1430	Not Present	1043	257	0.73	0.18	0.80	1300
Looking to caregiver	1430	Present	84	24	0.06	0.02	0.78	108
	1430	Not Present	1055	267	0.74	0.19	0.80	1322

Appendix F – Frequency of measures for Experiment 1 for younger and older children



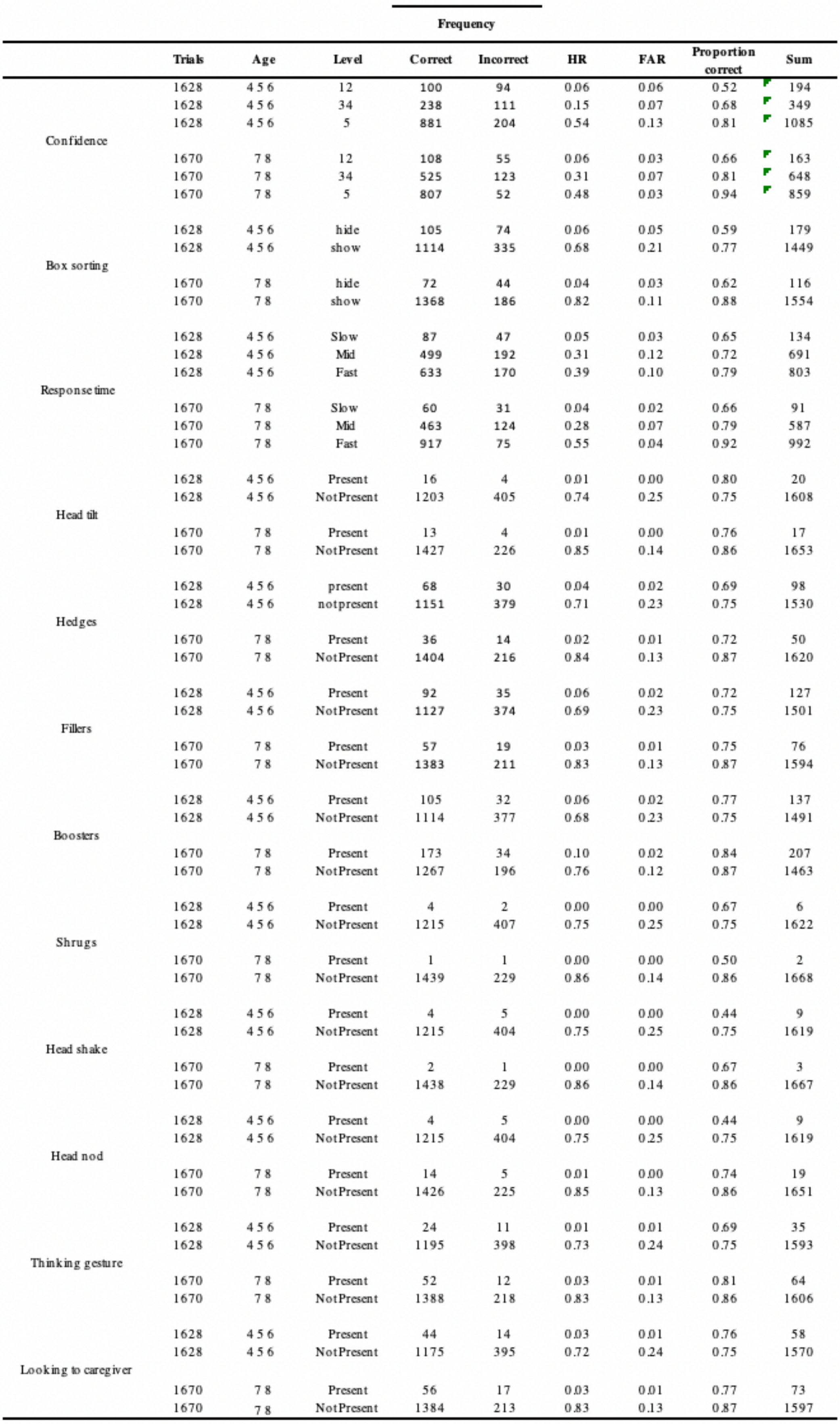
				Frequency					
	Trials	Age	Level	Correct	Incorrect	HR	FAR	Proportion correct	Sum
Confidence	841	4 5 6	12	37	29	0.04	0.03	0.56	66
	841	4 5 6	34	126	69	0.15	0.08	0.65	195
	841	4 5 6	5	485	95	0.58	0.11	0.84	580
	589	7 8	12	55	34	0.09	0.06	0.62	89
	589	7 8	34	184	55	0.31	0.09	0.77	239
	589	7 8	5	252	9	0.43	0.02	0.97	261
Box sorting	841	4 5 6	hide	50	18	0.06	0.02	0.74	68
	841	4 5 6	show	598	175	0.71	0.21	0.77	773
	589	7 8	hide	33	25	0.06	0.04	0.57	58
	589	7 8	show	458	73	0.78	0.12	0.86	531
Response time	841	4 5 6	Slow	68	40	0.08	0.05	0.63	108
	841	4 5 6	Mid	263	87	0.31	0.10	0.75	350
	841	4 5 6	Fast	317	66	0.38	0.08	0.83	383
	589	7 8	Slow	43	20	0.07	0.03	0.68	63
	589	7 8	Mid	148	46	0.25	0.08	0.76	194
	589	7 8	Fast	300	32	0.51	0.05	0.90	332
Head tilt	841	4 5 6	Present	32	15	0.04	0.02	0.68	47
	841	4 5 6	NotPresent	616	178	0.73	0.21	0.78	794
	589	7 8	Present	25	16	0.04	0.03	0.61	41
	589	7 8	NotPresent	466	82	0.79	0.14	0.85	548
Hedges	841	4 5 6	present	55	33	0.07	0.04	0.63	88
	841	4 5 6	notpresent	593	160	0.71	0.19	0.79	753
	589	7 8	Present	40	13	0.07	0.02	0.75	53
	589	7 8	NotPresent	451	85	0.77	0.14	0.84	536
Fillers	841	4 5 6	Present	54	29	0.06	0.03	0.65	83
	841	4 5 6	NotPresent	594	164	0.71	0.20	0.78	758
	589	7 8	Present	30	11	0.05	0.02	0.73	41
	589	7 8	NotPresent	461	87	0.78	0.15	0.84	548
Boosters	841	4 5 6	Present	107	17	0.13	0.02	0.86	124
	841	4 5 6	NotPresent	541	176	0.64	0.21	0.75	717
	589	7 8	Present	52	3	0.09	0.01	0.95	55
	589	7 8	NotPresent	439	95	0.75	0.16	0.82	534
Shrugs	841	4 5 6	Present	3	1	0.00	0.00	0.75	4
	841	4 5 6	NotPresent	679	198	0.81	0.24	0.77	877
	589	7 8	Present	6	0	0.01	0.00	1.00	6
	589	7 8	NotPresent	451	92	0.77	0.16	0.83	543
Head shake	841	4 5 6	Present	9	4	0.01	0.00	0.69	13
	841	4 5 6	NotPresent	673	195	0.80	0.23	0.78	868
	589	7 8	Present	3	0	0.01	0.00	1.00	3
	589	7 8	NotPresent	454	92	0.77	0.16	0.83	546
Head nod	841	4 5 6	Present	12	6	0.01	0.01	0.67	18
	841	4 5 6	NotPresent	670	193	0.80	0.23	0.78	863
	589	7 8	Present	3	0	0.01	0.00	1.00	3
	589	7 8	NotPresent	454	92	0.77	0.16	0.83	546
Thinking gesture	841	4 5 6	Present	69	23	0.08	0.03	0.75	92
	841	4 5 6	NotPresent	613	176	0.73	0.21	0.78	789
	589	7 8	Present	27	11	0.05	0.02	0.71	38
	589	7 8	NotPresent	430	81	0.73	0.14	0.84	511
Looking to caregiver	841	4 5 6	Present	42	14	0.05	0.02	0.75	56
	841	4 5 6	NotPresent	640	185	0.76	0.22	0.78	825
	589	7 8	Present	42	10	0.07	0.02	0.81	52
	589	7 8	NotPresent	415	82	0.70	0.14	0.84	497

Appendix G – Frequency of measures for Experiment 2 across all ages and trials



			Frequency					
	Trials	**Level**	**Correct**	**Incorrect**	**HR**	**FAR**	**Proportion correct**	**Sum**
Confidence	3298	1 2	208	149	0.06	0.05	0.58	357
	3298	3 4	763	234	0.23	0.07	0.77	997
	3298	5	1688	256	0.51	0.08	0.87	1944
Box sorting	3298	Hide	177	118	0.05	0.04	0.60	295
	3298	Show	2482	521	0.75	0.16	0.83	3003
Response time	3298	Slow	147	78	0.04	0.02	0.65	225
	3298	Mid	962	316	0.29	0.10	0.75	1278
	3298	Fast	1550	245	0.47	0.07	0.86	1795
Head tilt	3298	Present	29	8	0.01	0.00	0.78	37
	3298	Not Present	2630	631	0.80	0.19	0.81	3261
Hedges	3298	Present	104	44	0.03	0.01	0.70	148
	3298	Not Present	2555	595	0.77	0.18	0.81	3150
Fillers	3298	Present	149	54	0.05	0.02	0.73	203
	3298	Not Present	2510	585	0.76	0.18	0.81	3095
Boosters	3298	Present	278	66	0.08	0.02	0.81	344
	3298	Not Present	2381	573	0.72	0.17	0.81	2954
Shrugs	3298	Present	5	3	0.00	0.00	0.63	8
	3298	Not Present	2654	636	0.80	0.19	0.81	3290
Head shake	3298	Present	6	6	0.00	0.00	0.50	12
	3298	Not Present	2653	633	0.80	0.19	0.81	3286
Head nod	3298	Present	18	10	0.01	0.00	0.64	28
	3298	Not Present	2641	629	0.80	0.19	0.81	3270
Thinking gesture	3298	Present	76	23	0.02	0.01	0.77	99
	3298	Not Present	2583	616	0.78	0.19	0.81	3199
Looking to caregiver	3298	Present	100	31	0.03	0.01	0.76	131
	3298	Not Present	2559	608	0.78	0.18	0.81	3167

Appendix H – Frequency of measures for Experiment 1 for younger and older children



				Frequency					
	Trials	Age	Level	Correct	Incorrect	HR	FAR	Proportion correct	Sum
Confidence	1628	4 5 6	12	100	94	0.06	0.06	0.52	194
	1628	4 5 6	34	238	111	0.15	0.07	0.68	349
	1628	4 5 6	5	881	204	0.54	0.13	0.81	1085
	1670	7 8	12	108	55	0.06	0.03	0.66	163
	1670	7 8	34	525	123	0.31	0.07	0.81	648
	1670	7 8	5	807	52	0.48	0.03	0.94	859
Box sorting	1628	4 5 6	hide	105	74	0.06	0.05	0.59	179
	1628	4 5 6	show	1114	335	0.68	0.21	0.77	1449
	1670	7 8	hide	72	44	0.04	0.03	0.62	116
	1670	7 8	show	1368	186	0.82	0.11	0.88	1554
Response time	1628	4 5 6	Slow	87	47	0.05	0.03	0.65	134
	1628	4 5 6	Mid	499	192	0.31	0.12	0.72	691
	1628	4 5 6	Fast	633	170	0.39	0.10	0.79	803
	1670	7 8	Slow	60	31	0.04	0.02	0.66	91
	1670	7 8	Mid	463	124	0.28	0.07	0.79	587
	1670	7 8	Fast	917	75	0.55	0.04	0.92	992
Head tilt	1628	4 5 6	Present	16	4	0.01	0.00	0.80	20
	1628	4 5 6	NotPresent	1203	405	0.74	0.25	0.75	1608
	1670	7 8	Present	13	4	0.01	0.00	0.76	17
	1670	7 8	NotPresent	1427	226	0.85	0.14	0.86	1653
Hedges	1628	4 5 6	present	68	30	0.04	0.02	0.69	98
	1628	4 5 6	notpresent	1151	379	0.71	0.23	0.75	1530
	1670	7 8	Present	36	14	0.02	0.01	0.72	50
	1670	7 8	NotPresent	1404	216	0.84	0.13	0.87	1620
Fillers	1628	4 5 6	Present	92	35	0.06	0.02	0.72	127
	1628	4 5 6	NotPresent	1127	374	0.69	0.23	0.75	1501
	1670	7 8	Present	57	19	0.03	0.01	0.75	76
	1670	7 8	NotPresent	1383	211	0.83	0.13	0.87	1594
Boosters	1628	4 5 6	Present	105	32	0.06	0.02	0.77	137
	1628	4 5 6	NotPresent	1114	377	0.68	0.23	0.75	1491
	1670	7 8	Present	173	34	0.10	0.02	0.84	207
	1670	7 8	NotPresent	1267	196	0.76	0.12	0.87	1463
Shrugs	1628	4 5 6	Present	4	2	0.00	0.00	0.67	6
	1628	4 5 6	NotPresent	1215	407	0.75	0.25	0.75	1622
	1670	7 8	Present	1	1	0.00	0.00	0.50	2
	1670	7 8	NotPresent	1439	229	0.86	0.14	0.86	1668
Head shake	1628	4 5 6	Present	4	5	0.00	0.00	0.44	9
	1628	4 5 6	NotPresent	1215	404	0.75	0.25	0.75	1619
	1670	7 8	Present	2	1	0.00	0.00	0.67	3
	1670	7 8	NotPresent	1438	229	0.86	0.14	0.86	1667
Head nod	1628	4 5 6	Present	4	5	0.00	0.00	0.44	9
	1628	4 5 6	NotPresent	1215	404	0.75	0.25	0.75	1619
	1670	7 8	Present	14	5	0.01	0.00	0.74	19
	1670	7 8	NotPresent	1426	225	0.85	0.13	0.86	1651
Thinking gesture	1628	4 5 6	Present	24	11	0.01	0.01	0.69	35
	1628	4 5 6	NotPresent	1195	398	0.73	0.24	0.75	1593
	1670	7 8	Present	52	12	0.03	0.01	0.81	64
	1670	7 8	NotPresent	1388	218	0.83	0.13	0.86	1606
Looking to caregiver	1628	4 5 6	Present	44	14	0.03	0.01	0.76	58
	1628	4 5 6	NotPresent	1175	395	0.72	0.24	0.75	1570
	1670	7 8	Present	56	17	0.03	0.01	0.77	73
	1670	7 8	NotPresent	1384	213	0.83	0.13	0.87	1597