"BECAUSE SHE IS A KNOW-IT-ALL": SCHOOL-AGED CHILDREN'S UNDERSTANDING OF HESITANCE CALIBRATION

A Master's Thesis

by

ONUR SUNAY

Department of Psychology

İhsan Doğramacı Bilkent University

Ankara

July 2023

To those who died alone in the Kahramanmaraş Earthquake and their loved ones,

with hope without optimism

“BECAUSE SHE IS A KNOW-IT-ALL”: SCHOOL-AGED CHILDREN’S UNDERSTANDING OF CALIBRATION FOR HESITANT INFORMANTS

The Graduate School of Economics and Social Sciences
of
İhsan Doğramacı Bilkent University

by

ONUR SUNAY

In Partial Fulfillments of the Requirements for the Program of
MASTER OF ARTS IN PSYCHOLOGY

THE DEPARTMANT OF
PSYCHOLOGY
İHSAN DOĞRAMACI BILKENT UNIVERSITY
ANKARA

JULY 2023

“BECAUSE SHE IS A KNOW-IT-ALL”: SCHOOL-AGED CHILDREN’S UNDERSTANDING OF CALIBRATION FOR HESITANT INFORMANTS
By Onur Sunay

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Psychology.

……

Jedediah W. P. Allen

Advisor

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Psychology.

……

Hande Ilgaz

Examining Committee Member

I certify that I have read this thesis and have found that it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Arts in Psychology.

……

Dilay Z. Karadöller

Examining Committee Member

Approval of the Graduate School of Economics and Social Sciences

……

Refet Soykan Gürkaynak

Director

ABSTRACT

"BECAUSE SHE IS A KNOW-IT-ALL": SCHOOL-AGED CHILDREN'S UNDERSTANDING OF CALIBRATION FOR HESITANT INFORMANTS

Sunay, Onur

MA, Department of Psychology

Advisor: Asst. Prof. Jedediah W.P. Allen

Calibration refers to the extent to which one's confidence predicts their accuracy. Accordingly; someone accurate and confident, and someone inaccurate and hesitant are well-calibrated; and someone inaccurate and confident, and someone accurate but hesitant are poorly calibrated. Although there is evidence of adults' calibration understanding, children do not have a complete understanding of calibration. The current study aimed to investigate children's calibration understanding better. To that end, 7-, 9-, and 11-year-old children were tested on three calibration tasks with informants that included the inaccurate and hesitant informant. The tasks included explicit and implicit measures of calibration. The results showed that children performed similarly across all ages, but there were differences in how children performed between different tasks. Also, accuracy had more influence on children's judgments for who was a reliable informant than confidence. Third, more children passed the implicit calibration task but failed the explicit one than vice versa. Lastly, children's calibration understanding was not related to their executive function (EF) abilities. These results suggest that calibration is a complex ability influenced by social situations. The role situations play and how they might be used as a broader framework to explain calibration are highlighted in the discussion. EF and other

cognitive abilities that might be related to calibration understanding are also discussed.

Keywords: calibration, accuracy, confidence, selective trust, credibility

ÖZET

“ÇÜNKÜ O UKALA”: OKUL ÇAĞINDAKİ ÇOCUKLARIN TEREDDÜTLÜ KAYNAKLARA YÖNELİK KALİBRASYON ANLAYIŞI

Sunay, Onur
Yüksek Lisans, Psikoloji Bölümü
Tez Danışmanı: Dr. Öğretim Üyesi Jedediah W.P. Allen

Kalibrasyon (Doğruluk-Tereddüt Uyumu), kişinin kendine olan güveninin doğruluğunu ne ölçüde öngördüğünü ifade eder. Buna göre; doğru ve kendine güvenen biri ile yanlış ve tereddütlü biri sağlam uyumlu; yanlış ve kendine güvenen biri ile doğru ama tereddütlü biri ise zayıf uyumludur. Yetişkinlerin kalibrasyon anlayışına dair bulgular olmasına rağmen çocukların kalibrasyon konusunda tam bir anlayışa sahip değildir. Bu çalışma, çocukların kalibrasyon anlayışını daha iyi araştırmayı amaçlamıştır. Bu amaçla, 7, 9 ve 11 yaşındaki çocuklar, yanlış ve tereddütlü kaynakları içeren üç kalibrasyon görevinde test edilmiştir. Görevler kalibrasyonun açık ve örtük ölçümlerini içeriyordu. Sonuçlar, çocukların üç yaş grubunda da benzer performansının olduğunu ancak çocukların farklı görevler arasında nasıl performans gösterdikleri konusunda farklılıklar olduğunu göstermiştir. Ayrıca doğruluk, çocukların kimin güvenilir bir kaynak olduğuna ilişkin yargıları üzerinde güvenden daha fazla etkiye sahipti. Üçüncü olarak, daha fazla çocuk örtük kalibrasyon görevinden geçmiş ancak açık görevde başarısız olmuştur. Son olarak,

çocukların kalibrasyon anlayışı ile yürütücü işlev yetenekleriyle arasında bir ilişki bulunmamıştır. Bu sonuçlar, kalibrasyonun sosyal durumlardan etkilenen karmaşık bir yetenek olduğunu göstermektedir. Sosyal durumların sahip olduğu rol ve kalibrasyonu açıklamak için daha geniş bir çerçeve olarak nasıl kullanılabilecekleri tartışmada vurgulanmaktadır. Kalibrasyon anlayışı ile ilişkili olabilecek yürütücü işlevler ve diğer bilişsel yetenekler de tartışılmaktadır.

Anahtar Kelimeler: kalibrasyon, doğruluk, kendine güven, seçici güven, güvenilirlik

ACKNOWLEDGEMENTS

I should first thank my advisor Jed Allen and Hande Ilgaz. I can only imagine the headaches they had trying to understand me and help me improve my ideas. They helped me improve as a thinker, and even more as a writer. They raised a student confident enough to say “No, that’s not how Piaget thinks,” which is easier said than done.

Although I wrote and defended this thesis individually, it is group work. I want to thank Atabey Yurtsev, Beyza Übeyli, Beyza Tekbaş, Deniz Coşkun, Ece Yüksel, İpek Salman, İrem Önal, İzem Çelik, Parisa Karimpour, Zafer Coşkun for their interest in the project and support they provided at different stages of the thesis. They helped me stay motivated to pursue the project.

I would like to thank Ece, İdil Ece, Meryem, Nisan, Rana, and Şevval for their time in acting for the informant videos that I used in the current study. Whether they know it or not, their support helped me reach one of the most important milestones of my thesis. Among these actors, I owe special thanks to Ece Tuğlacı. She did not only

help me find actors, but she also has been my thesis-comrade throughout this year. Ece, please keep sending memes, and may dad jokes be with you.

I thank the residents of Graduate Suite II, Gaye Aşkın, Gökçe Şahin, and Nihal Gökçe. They were the greatest “colleagues” I could have during this year. We may have chatted where we should have studied, but for good reasons. I will miss the atmosphere we created. I especially thank Gökçe Şahin for the comradery in a second project and Eda Önoğlu Yıldırım for her advice in reaching schools.

I also would like to thank Ahmet Gönüllü, Aycan Funda, Burak Doğancıoğlu, Büşra Göçer, Cemali Arıcı, Ece Gökçe, and Zeynep Sunay for their long and enduring friendship. “They treat me so good I begin to feel special … I said they treat me so well I feel special … Thank you. Thank you.”

Lastly, I would like to thank my parents for their support and encouragement by all means. I was lucky to have such enablers in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

"The true direction of the development of thinking is not from the individual to the social, but from the social to the individual."
– Lev S. Vygotsky

While most animals tend to learn from individual and observational experiences, humans are massively social in their learning. A core aspect of social learning is language (Heyes, 2012). Language enables humans to learn indirectly through verbal discourse or from written sources. The indirect nature of this learning introduces complexities to the learning situation. Specifically, social learning situations involving testimony have a problem. If the source of knowledge lacks credibility, it is reasonable that someone chooses not to believe them.

In learning from others through testimony, people can decide and act based on several factors. Accuracy and confidence are two cues that determine credibility in that accurate and confident people are more favorable sources of information, i.e.,

informants. However, for research, accuracy and confidence alone are not the ideal units to analyze in isolation (Tenney et al., 2008). For instance, someone who was previously inaccurate but confident, or a "know-it-all," is not reliable despite the apparent confidence. This person may look charismatic and convincing but would lack credibility. The extent to which one's confidence predicts accuracy—coined in the literature as *calibration*—is a complex and more comprehensive unit to analyze. Within the scope of calibration, informants can be well-calibrated when they are either accurate (and confident) or inaccurate (and hesitant) (Tenney et al., 2008, 2011). It may be relatively easy for adults to have an understanding of calibration, but such understanding is an outcome of development. The current study investigates the extent to which school-aged children (7-, 9-, and 11-year-olds) understand the complex nature of calibration, specifically hesitance calibration (i.e., inaccurate and hesitant informants).

1.1. Calibration as a Form of Selective Trust

People learn from both first-hand experience and testimony: one can see the seagulls flying to the South, or one can be told that such is the case. Learning from testimony is possible only if one takes others' credibility for granted. Namely, one must trust others to learn from them. Testimony has been studied within the scope of epistemologists (e.g., Audi, 1997) and psychologists. Psychologists' work on testimony is multifaceted, spreading to different fields, including forensic psychology (Spore, 1982) and developmental psychology (Fusaro & Harris, 2013).

As studied in developmental psychology, *selective trust* research investigates factors related to children's preferences of whom to trust when there is conflicting

information between informants (Koenig et al., 2004). At a general level, selective trust concerns children's choice of informant. For instance, one should see a doctor in times of sickness and a mechanic when the car breaks down, but not vice versa. This example illustrates how expertise as an indicator of competence is a relevant feature of selective trust. Other interpersonal features such as being deceptive or disloyal also affect testimony (Koenig & McMyler, 2019; Tong et al., 2020).

The two cues involved in calibration (accuracy and confidence) are concurrently present and coupled. What implications does the intertwined nature of these cues have for calibration? What makes them intertwined in the first place? Accuracy and confidence are interlinked cues for two reasons. First, if accuracy is with respect to the content of a statement, confidence is part of the process of characterizing the content. Where accuracy is about what is being said, confidence is about how it is being said. That is, it is not possible to say something without saying it in a certain way. The second reason accuracy and confidence are intertwined is reflected in the core assumption of calibration: if confidence predicts accuracy, then they are relational, and the valence of one potentially changes the positive or negative impact of the other.

Accuracy and confidence map onto relatively epistemic and social aspects of testimony, respectively; they are also relatively objective and subjective (Brosseau-Liard et al., 2014). Accuracy is objective since the truth of any information constitutes accuracy. Where people's accuracy concerns whether they know something, their confidence is multifaceted. On the one hand, confidence indicates

credibility or persuasiveness (Tenney et al., 2007, 2008). On the other hand, confidence reflects the informants' self-monitoring. Given its relation to accuracy, confidence indirectly provides some grounds for whether someone is accurate. This is the core idea behind calibration.

Although accuracy and confidence are distinct and can sometimes be analyzed separately, calibration involves a whole greater than the two-component parts (i.e., calibration is constituted by the relationship between the two). When a person's accuracy or confidence changes, they become poorly calibrated if they were well-calibrated before, and vice versa. This is not because of a difference in one's accuracy or confidence alone but because of the relationship between some semantic/factual information and how an informant presents such information. Accordingly, four categories emerge based on people's accuracy and confidence. Well-calibrated people are[1]:

1. Accurate *and* Confident (Confidently Well-Calibrated): The Good Winner

2. Inaccurate *and* Hesitant (Hesitantly Well-Calibrated): The Good Loser

Poorly calibrated people are:

1. Accurate *but* Hesitant (Accurately Poorly Calibrated): The Skeptic or The Lucky

[1] Throughout the thesis, informant characteristics will be written openly and abbreviated. In the abbreviations, "A" stands for accurate, "I" stands for inaccurate, "C" stands for confident, and "H" stands for hesitant.

2. Inaccurate *but* Confident (Confidently Poorly Calibrated): The Know-It-All (overconfident)[2]

Of note, it is possible to be well-calibrated when accurate or *inaccurate*. Thus, calibration is not about who is right, at least not exclusively. Instead, calibration is about the extent to which confidence predicts accuracy to determine whether informants are credible. Credible informants may not always be correct, but they are still credible if they show that they know what they do not know. However, credibility is more straightforward when an informant is accurate. That is, recognizing someone's *negative credibility* (e.g., someone inaccurate and well-calibrated) is more difficult than their positive credibility (e.g., someone accurate and well-calibrated).

1.2. Is Calibration Well-Calibrated?

Calibration is related to selective trust, but it has been studied less. Research on calibration was pioneered in the social psychology literature and carried over to developmental psychology more recently. The first calibration studies targeted the

[2] One might ask if calling an inaccurate but confident person an overconfident "know-it-all" is justified. To illustrate the case, think of the standard change of location task. In this task, changes in reality take place unbeknownst to some people. We have our prior knowledge through our prior experiences. Any change in reality (e.g., changing the location of a letter) will be known through further experience (either first-hand or not). Those who do not "update" their knowledge will be unaware of some facts. These people might be confident in their answers. This, however, does not make them overconfident fools. This is a point that requires elaboration.
There is a difference between giving a wrong answer to a question that you do not know the answer to as opposed to one that you have a justified answer to if your prior knowledge were still true. That is, there is a difference between saying "The purse is in the kitchen" and "All mammals fly." The former is epistemic, such that the purse could be replaced unbeknownst to the informant. The other sentence is semantic, such that it is accepted as an objective fact regardless of conditional changes.
Is this distinction necessary? Yes. If there were no such distinction, the person who has false beliefs in standard ToM tasks would be in the same position as the inaccurate one in the trust task. There is a reason why the person in the ToM task says something wrong, but there is no good reason for the inaccurate person in the trust task. Therefore, the latter should not be attributed as credible.

confidence heuristic in the literature and showed that people do not show a simple confidence heuristic (i.e., high confidence is high reliability). The general findings were that people are rather sensitive to informant calibration when responding to witness cases in an accident (Tenney et al., 2007, 2008). In these studies, two informants provide different information about an accident. Informants displayed different accuracy and confidence characteristics. When the only available cue was confidence, the participants preferred the confident informants. When accuracy was also revealed, they switched to the well-calibrated hesitant informant (Inaccurate-Hesitant) from the poorly calibrated confident informant (Inaccurate-Confident). When a further explanation was given for why the poorly calibrated confident informant was inaccurate (e.g., they had a mistaken belief), participants reversed their preferences to their initial choice (i.e., they excused the poor calibration). Further, the "poorly calibrated" confident informant was now also rated to be as honest and credible as the well-calibrated hesitant informant (Tenney et al., 2008). These studies showed that people are sensitive to confidence when calibration information is unavailable and switch to calibration once presented (i.e., once accuracy is introduced). Further, people can change their responses depending on situational constraints (i.e., reasons for poor calibration).

Tenney and colleagues showed how flexible people can be in their calibration judgments. They also showed how small changes can impact people's decision-making processes toward whom to trust. Further research showed that adults do not make decisions based on calibration when they were under cognitive load (instead, they used confidence alone). This suggests that executive functions (EF) are a relevant component for calibration (Tenney et al., 2011, Experiment 4).

Tenney and colleagues (2011) also tested children in an adaptation of their original paradigm and found that 5- and 6-year-old children do not understand calibration. Although this adapted paradigm yielded results similar to a selective trust (Experiment 1b), children (incorrectly) preferred the poorly calibrated confident informant (IC) over the well-calibrated hesitant informant (IH; Experiment 1a). Another experiment showed that when children heard two confident informants testifying for two pieces of information, children preferred the well-calibrated confident informant (AC) instead of the poorly calibrated confident informant (IC; Experiment 3). This means that children preferred the accurate informant who was also well-calibrated and acted based on accuracy but not calibration when both informants were confident.

Lastly, when children were shown a confident and a hesitant informant, they initially preferred the confident informant. When both informants were shown to be accurate, children did not choose one informant over the other (Experiment 2). That is, they did not show calibration understanding because they did not differentiate between the well-calibrated and the poorly calibrated accurate informants. In short, 5- and 6-year-old children are more sensitive to accuracy than confidence, and they do not have calibration understanding (Tenney et al., 2011). In parallel, Brousseau-Liard and colleagues (2014) showed that when children saw a poorly calibrated accurate informant (AH) and a poorly calibrated confident informant (IC), they started to make endorsements based on accuracy rather than confidence beginning around four years of age.

More recently, Birch and colleagues (2020) challenged Tenney and colleagues (2011) and suggested that children calibrate for confidence as early as five years of age, but even 8-year-old children do not have hesitance calibration (i.e., the hesitant informant can be well-calibrated). Correspondingly, *confidence calibration* and *hesitance calibration* highlight calibration understanding for accurate and confident informants (AC) and inaccurate and hesitant informants (IH). Birch and colleagues motivated their claim by measuring children's preferences and smartness judgments for informants. Specifically, Birch and colleagues (2020) showed that children prefer to learn from a well-calibrated accurate informant (AC) over a poorly calibrated accurate informant (AH; Experiment 1 informed condition). In this condition, informants are either well-calibrated through confidence or poorly calibrated through hesitance, but all informants are accurate. However, this creates a confound in that the calibration of informants varies in the same direction with their accuracy or their confidence. This means that children's preference for a well-calibrated informant over a poorly calibrated one does not necessarily mean children understand calibration. They may just be choosing based on accuracy or confidence alone because informants' calibration changes in the same direction as their accuracy or confidence. That is, when a well-calibrated accurate informant becomes inaccurate, the informant also becomes poorly calibrated. Being inaccurate and poorly calibrated are both not preferable. A negative difference between their accuracy leads to a negative difference in their calibration. In this case, the different displays of confidence (or accuracy) correspond to being calibrated (i.e., confidence makes calibration well, and hesitance makes calibration poor).

The same applies to a condition with a well-calibrated confident informant (AC) and a poorly calibrated confident informant (IC): a preference based on calibration would be the same as a preference based on accuracy. Birch and colleagues showed that children endorsed the well-calibrated/accurate informant more than the poorly calibrated/inaccurate informant when both informants were confident (Experiment 2). Birch and colleagues (2020) interpreted these findings to mean that children have an understanding of calibration; however, they also equivocate on the uses of children "preferring the well-calibrated informant" and "understanding calibration." This highlights a confound in the design.

As shown in Figure 1, hesitance and calibration differences are in the same direction between a well-calibrated accurate informant (AC) and a poorly calibrated hesitant informant (AH; as in Experiment 1, Informed condition), and between a well-calibrated accurate informant (AC) and a poorly calibrated confident informant (IC; as in Experiment 2). That is, a negative difference in confidence (2a) leads to a negative difference in calibration (2a). Similarly, a negative difference in accuracy (2b) leads to a negative difference in calibration (2b). Consequently, whether any difference in children's preferences results from calibration or accuracy cannot be distinguished with such a design.

One possible solution to the design problem is to focus on hesitant informants while keeping accuracy constant. An inaccurate and hesitant informant (IH) is well-

Figure 1. Relation of Informants in Confidence Calibration

Informant	Accuracy		Confidence		Calibration	
Informant 1a	Accurate	(+)	Confident	(+)	Well-Calibrated	(+)
Informant 2a	Accurate	(+)	Hesitant	(-)	Poorly Calibrated	(-)
Informant 2b	Inaccurate	(-)	Confident	(+)	Poorly Calibrated	(-)

Note. Informants 1a-2a: A negative difference in confidence leads to a negative difference in calibration. Informants 1a-2b: A negative difference in accuracy leads to a negative difference in calibration.

calibrated. In contrast, a confident but inaccurate informant (IC) indicates poor calibration. Similarly, being accurate when hesitant (AH) means poor. As shown in Figure 2, a positive difference in accuracy (2a) leads to a negative difference in calibration. Similarly, a positive change in confidence (2b) leads to a negative change in calibration. Unlike confidence calibration, for hesitance calibration, positive differences in accuracy or confidence lead to negative differences in calibration (i.e., the confound between accuracy and calibration is removed). Note that Tenney and colleagues' (2011) null results for children with an understanding of calibration are from informant pairs (1b-2a and 1b-2b) in Fig. 2 (i.e., no confound design). On the other hand, the results of calibration understanding from Birch and colleagues have had informants, as illustrated in Fig. 1 (i.e., a confounded design).

Figure 2. Relation of Informants in Hesitance Calibration

Informant	Accuracy		Confidence		Calibration	
Informant 1b	Inaccurate	(-)	Hesitant	(-)	Well-Calibrated	(+)
Informant 2a	Accurate	(+)	Hesitant	(-)	Poorly Calibrated	(-)
Informant 2b	Inaccurate	(-)	Confident	(+)	Poorly Calibrated	(-)

Note. Informants 1b-2a: A positive difference in accuracy leads to a negative difference in calibration. Informants 1b-2b: A positive difference in confidence leads to a negative difference in calibration.

However, in a non-confound Experiment 3, Birch and colleagues (2020) showed that even 8-year-old children did not prefer the well-calibrated hesitant informant (IH) over the poorly calibrated accurate informant (AH). In another experiment, when the well-calibrated hesitant and the poorly calibrated confident informant were in the same condition, 5-year-old children preferred the well-calibrated informant (IH) over the poorly calibrated informant (IC; Birch et al. 2020, Experiment 1, Uninformed Condition). The last finding suggests that children understand hesitance calibration, which is interesting because Birch and colleagues (2020) do not make this inference. If anything, this shows that studying calibration requires extra caution. Calibration as a sub-field of selective trust is not yet well-calibrated but is still in progress.

1.3. Towards a Better-Calibrated Calibration Study

The current study aims to expand the work on calibration. To that end, the current study engages critically with Tenney and colleagues (2011) and Birch and colleagues

(2020). The first decision is informant choice. Which informant pairs are the most relevant? The current study includes three conditions to better explore the developmental trajectory of calibration understanding. The first condition includes two hesitant informants (1b and 2a, Fig. 2), one of which is well-calibrated (IH), and the other informant is poorly calibrated (AH). The second condition has two inaccurate informants (1b and 2b), one of which is well-calibrated (IH) and the other is poorly calibrated (AC). The third condition includes the two well-calibrated informants (1a and 1b), whether they are accurate (AC) or inaccurate (IH). In short, a well-calibrated hesitant informant (IH) is present in each condition and is paired with one of the other three informant types. The first two conditions are for hesitance calibration and were selected based on previous literature. However, as argued above, confidence calibration does not allow researchers to distinguish "preferring a well-calibrated informant" from "having an understanding of calibration" because preferences could be based on accuracy alone. Condition 3 is novel as no previous research looked at potential differences between the two ways of being well-calibrated (see Table 1 for a detailed description of informant pairs in the current study).

The current study examines calibration understanding in school-aged children. To that end, 7-, 9-, and 11-year-old children were sampled. The age range is above that of Tenney and colleagues' sample who failed in the calibration tasks. The age range also includes and exceeds Birch and colleagues' sample, who found mixed results for calibration understanding. Thus, the current study can yield implications regarding different phases of calibration understanding. Age 9 is potentially a milestone for children's calibration understanding because children show an understanding of

situations where experts may not know a fact because it is something difficult to know (i.e., virtuous ignorance) at that age (Kominsky et al., 2016).

The current study measures executive functions (EF) as an indicator of cognitive flexibility and working memory in children's development. This characterization is motivated by Tenney and colleagues' (2011) findings that showed adults do not calibrate under cognitive load. A modified version of the Dimension Change Card Sort task (DCCS) measures EF (Bock et al., 2015). The modified DCCS task (DCCSm) measures cognitive flexibility with higher working memory demands than the original DCCS.

Lastly, the current study aims to expand the type of questions asked to children in calibration studies. Ask and endorse questions are common in selective trust research and used in calibration research. Credibility judgments (Tenney et al., 2011) and smartness judgments (Birch et al., 2020) are examples of further questions that may be asked. In the current study, children were asked how trustworthy they thought the informants were (trustworthiness question), and how informants would act in an alternative situation (hypothetical thinking question). Trustworthiness and Hypothetical Thinking questions enabled the current study to explore children's understanding of calibration in terms of an explicit and an implicit measure, respectively. Further details on these questions are discussed in the Methods section.

Table 1. A Summary of Conditions in the current study

	Informant 1 (IH)				Informant 2 (Other)			
	Informant	**Confidence**	**Accuracy**	**Calibration**	**Informant**	**Confidence**	**Accuracy**	**Calibration**
Condition 1	**IH**	**Hesitant**	**Inaccurate**	**Well-Calibrated**	**AH**	**Hesitant**	**Accurate**	**Poorly Calibrated**
Condition 2	**IH**	**Hesitant**	**Inaccurate**	**Well-Calibrated**	**IC**	**Confident**	**Inaccurate**	**Poorly Calibrated**
Condition 3	**IH**	**Hesitant**	**Inaccurate**	**Well-Calibrated**	**AC**	**Confident**	**Accurate**	**Well-Calibrated**

Note. The red font (underlined) highlights negative informant characteristics; the green font highlights positive informant characteristics.

1.4. The Current Study

The aim of the current study is to investigate the development of children's understanding of calibration, with a particular focus on hesitance calibration (i.e., the well-calibrated hesitant informant). Thus, the question is how and when children come to understand calibration. A complete answer to this question should include additional measures that may tap into both explicit and more implicit processes. To that end, the current study tests the hypotheses listed below:

Research Question 1: When do children develop an understanding of calibration?

Hypothesis 1: Children develop some understanding of calibration starting at 9- and improving by age 11.

Prediction 1: Older children *endorse* the well-calibrated hesitant informant (IH) more when she is paired with a poorly calibrated informant (Conditions 1 & 2).

Prediction 2: Older children judge the well-calibrated hesitant informants as relatively more *trustworthy* and poorly calibrated informants as less *trustworthy*.

Prediction 3: Older children perform better on all *hypothetical thinking* questions.

Research Question 2: Which of the two components of calibration has a stronger effect: accuracy or confidence?

Hypothesis 1: Accuracy (the objective cue of calibration) is more salient for calibration than confidence (the subjective cue).

Prediction 1: Children *endorse* informants based on accuracy more than confidence (AH over IC).

Prediction 2: Children rate the poorly calibrated accurate informant (AH) in Condition 1 as more *trustworthy* than the poorly calibrated confident informant (IC) in Condition 2.

Prediction 3: Children perform better on the *hypothetical thinking* question for Condition 2 than for Condition 1.

Research Question 3: Is calibration manifested differently when presented implicitly or explicitly?

Hypothesis 1: An implicit understanding of calibration precedes an explicit understanding of calibration.

Prediction 1: More children pass the hypothetical thinking question but fail the trustworthiness question than vice versa.

Research Question 4: Is EF a relevant component for children's understanding of calibration?

Prediction 1: Children's performance on DCCSm and calibration tasks are positively correlated.

CHAPTER II

METHODS

2.1. Participants

Ninety-five 7-, 9-, and 11-year-old children participated in the current study. Children were recruited from public schools in the Çankaya district of Ankara, the capital of Turkey, and they were mainly from upper-middle SES families. Parent reports indicated that 74.7% of the mothers had a university degree or above. A 7-point income scale showed that 70.50% of children came from families who reported their family income as four or above. One 7-year-old child was excluded from analyses due to poor Turkish comprehension. The final sample consisted of 33 7-year-old children (M = 89.25, SD = 3.71, 82-96 months), 29 9-year-old children (M = 114.04, SD = 3.99, 104-121 months), and 32 11-year-old children (M = 129.23, SD = 3.00, 125-136 months). Note that the oldest age group has a mean age toward the end of being 10.

2.2. Materials and Procedure

The current study was approved by the Ethics process of Bilkent University and the Turkish Ministry of National Education. After approaching schools and getting their permission to run the study, consent forms were distributed to parents. Of those parents who consented to their children's participation in the current study, children were approached via their teachers for their verbal assent. No child participated in the study if they did not want to or if the consent was not completed by parents. Children were tested in the library of their schools. They were tested individually by an experimenter and typically with a coder. Each testing session lasted 35-40 minutes. When the experimenter was alone, the testing session was audio-recorded. No audio was recorded when a coder was present during testing.

The three conditions tested in the current study consisted of three pairs, Condition 1: well-calibrated hesitant (IH) and poorly calibrated accurate (AH), Condition 2: well-calibrated hesitant (IH) and poorly calibrated confident (IC), and Condition 3: well-calibrated hesitant (IH) and well-calibrated confident (AC; see Table 1). Condition 3 was always presented last, and the order in which Conditions 1 and 2 were presented was counterbalanced.

At the start of testing, children were seated in front of a 12.5-inch laptop. They were then introduced to the general procedure with the opening: "We will play a trivia game with you. First, we will watch documentary videos of little-known animals. Then, we will watch others who played this game. We will talk about how well they played. After that, I will show you another animal. I will be asking you some

questions while we are playing the game. We will play this game three times. Lastly, we will play a card game." After all the "trivia games," the laptop was removed from the table to play DCCSm, the card game.

2.2.1. Endorsement Task: Induction Phase

2.2.1.1. Documentary Videos

Two design decisions were made in producing the videos that differ somewhat from the previous literature. These are the use of accuracy instead of knowledge access and object properties instead of object names. Unlike Birch et al.'s (2020) use of knowledge access, the current study manipulates accuracy as the epistemic cue. Given the developmental discrepancy between false-belief understanding (as an understanding of (in)accuracy) and knowledge access (Wellman & Liu, 2004), accuracy is more directly to the point of someone's credibility through calibration.

The current study uses object (i.e., animal) properties instead of names for the induction phase. This differs from most trust research, which uses the naming of everyday objects (e.g., the accurate informant calls a cup a cup, but the inaccurate one calls it a fork accuracy as a cue). This is mainly motivated by the fact that children in the current study are older. Older children may think informants are deliberately deceptive in naming everyday objects inaccurately.

For each condition, children watched a short documentary. The documentaries were taken from a series developed by the Turkish Radio and Television Corporation

(Turkey's public broadcaster). The videos are between 1.5 and 2.5 minutes in length. See Appendix A for sample snapshots from the documentaries.

Following the documentaries, children answered memory check questions to ensure they learned the relevant information that the informants would tell in the next step. If children did not respond correctly to a question, the experimenter replayed the part relevant to the question. No child saw a part more than twice. See Appendix B for each video's list of questions and their correct and incorrect answers.

2.2.1.2. Informant Clips

Children saw informants after memory questions. For each condition, children saw two informants with characteristics described in the introduction. One informant in each pair was always a well-calibrated hesitant informant (IH). In total, there were six informants. In Condition 1, the second informant was poorly calibrated and accurate (AH); in Condition 2, poorly calibrated and confident (IC); and in Condition 3, well-calibrated and confident (AC). No pair of informants had a notably different physical appearance. Informants in each condition were counter-balanced, such that each informant appeared on the left of the screen half the time. Each pair of informants answered questions for the same documentary.

Informant responses to questions were pre-recorded and were shown on the same laptop using Microsoft PowerPoint. During this phase, children were first introduced as players playing the same "trivia game." Children were explicitly told, "[The informants] may or may not have seen the documentary we have just seen. We do

not know." This was done to ensure that children do not think the informants are being funny, deceptive, or anything that disrupts the accuracy manipulation necessary for the current study.

Informants in each condition either wore a red or blue t-shirt. Because of the lighting during the shooting of informant clips, the blue t-shirt appeared black. Consequently, the blue t-shirt was treated as black. The t-shirt color (red or black) and informants' display order (whether one informant was shown first on the left of the screen or on the right of the screen) were counterbalanced for each pair.

For each question, children saw both informants on the screen at the same time. The question informants answered was also visible for each question. Children or the experimenter read the questions first. Then children were asked to answer the question. This served as a double-check for the memory question. Following that, children saw the informant clips of the conditions defined by the informants answering accurately or inaccurately, confidently or hesitantly. Confidence and hesitance were operationalized similarly to Birch and colleagues (2020). Confident informants spoke with a declarative tone, they pointed their fingers upwards, and they started their sentences with "I know this," "Aha," and "I know this too." Hesitant informants spoke with a soft tone, shrugged their shoulders, and started their sentences with "I guess," "Hmm," and "I think/suppose." See Appendix C for snapshots from the informant clips. Each question was presented one at a time, and informants responded to questions one by one.

2.2.1.3. Manipulation Check

After seeing all four questions, children were asked if Informant A and Informant B were accurate or inaccurate, and confident or hesitant separately for each informant. Prior to the manipulation check, some children were scaffolded during informant clips to ensure they understood the manipulation. For instance, they were told, "Someone who says 'Hmm, I guess' is not sure of herself." Children who failed the manipulation check questions were excluded from the analysis.

2.2.2. Endorsement Task: Test Phase

2.2.2.1. Endorsement

After the manipulation check, children were asked the endorsement question. Children were shown a picture of a novel animal. Animals were chosen as stimuli to stay in the same domain for induction and testing. Children were asked if they knew the animal. If children only claimed to know the animal but did not actually know it, they were told it looked like the one shown in the photograph but was actually some other animal. If children did not know the animal, the task continued. If children knew the animal, additional backup pictures were shown. Next, the accurate informants said the actual name of the animals, and the inaccurate informants said similar fake names. Some children asked the animals' names to the experimenter. In these cases, the experimenter said he could not say it now and provided children with the actual names at the end of the experiment.

For each condition, children were told what Informant 1 and Informant 2 think the novel animal is by the experimenter. The experimenter told the informants'

responses without intonation. This was done to eliminate confidence during the test phase[3]. Informants' confidence was unknown to children. The test phase was explicitly designed so that children did not see matches or mismatches between induction and testing. Some children asked whether informants were confident or hesitant when answering these questions. When this was the case, the experimenter said that the informants were asked to write down their answers. Therefore, from the child's perspective, the experimenter could not have known the answer to the children's question either.

The endorsement question directed to children was, "What do you think this animal is?". If the children did not respond within 10 seconds, the experimenter asked, "Do you think it is [Answer 1], [Answer 2], or does it have another name?"[4] Children's responses were coded as IH's answer, other informant's answer, or neither. Following the children's responses, the experimenter asked why they thought that. Children's explanations were not analyzed in the current study.

2.2.2.2. Trustworthiness Question

The trustworthiness question is considered a more direct measure of calibration than Hypothetical Thinking. In the trustworthiness question, children were asked how

[3] The current study deviates from Birch et al. (2020) in asking whom children endorse. In that study, children saw a confident and a hesitant informant in the induction stage. In testing, however, both informants were confident. A sudden change in hesitant—but not confident—informants' shift to confidence in testing creates a potential confound. The experimenter told the informants' responses to overcome this potential confound in the current study.

[4] The endorsement question in the current study is open-ended. Unlike Birch et al. (2020), children were given the option to prefer neither informant. Since the current study focuses on hesitance calibration and has a pair of two inaccurate informants, the option not to prefer any informant is beyond superficial.

trustworthy they thought the informants were. They rated informants' trustworthiness on a 1 to 7 scale using cartoon faces and short texts, not numbers. For instance, children saw a red upset face doing a thumbs-down and (were) read "not Trustworthy at all" for a score of 1. See Appendix E for the full scale of trustworthiness.

2.2.2.3. Hypothetical Thinking Question

The Hypothetical Thinking question is novel to calibration and trust research. It is being used in the current study as an implicit measure of calibration understanding. The assumption motivating this question is the *consistency of calibration*: if informants are confident when accurate, they are expected to be hesitant when inaccurate, and vice versa.

The Hypothetical Thinking question started with a reminder of whether the informants were confident or hesitant in the informant clips. Then, children were asked whether informants would say something True or False about the child's favorite animal. The experimenter generally asked the children's favorite animal while walking to the testing room with the child. This early conversation served as a warm-up. When not asked at the beginning, the experimenter asked about the children's favorite animal right before the start of the Hypothetical Thinking question.

Children answered Hypothetical Thinking questions for each informant, but only responses for the well-calibrated informants were analyzed. Thus, four informants across the three conditions (three hesitant (IH), one confident (AC)) were included

for analyses. Children received 1 when they responded True for the well-calibrated hesitant informant (IH) and False for the well-calibrated confident informant (AC). Children received 0 when they responded False for the well-calibrated hesitant informant (IH) and True for the well-calibrated confident informant (AC). The full text of the question is available in Appendix D.

2.2.3. DCCSm

A modified version of the original DCCS task (Zelazo, 2006) was used to measure children's EF (Bock et al., 2015). The revised version has the same essential structure as the original but has higher working memory demands. The original DCCS has two simple rules: color and shape with two dimensions (red and blue for color, rabbit and boat for shape). The modified DCCS (DCCSm) has three rules with three dimensions. As used in the current study, color (red, blue, yellow), shape (circle, triangle, square), and number (two, four, six) were the variables changing between cards. There were three reference cards for children to sort under, three demonstration cards for Games 1-3, six testing cards for Games 1-3, three for demonstration, and twelve for testing in Game 4.

Children first played sorting games according to three rules, one at a time. The order in which rules were presented was counterbalanced. Children were first told the rule and shown a demonstration of each game. Then, children sorted six cards. When changing between rules, the experimenter said, "We are no longer playing [Game 1]. Now, we are going to play [Game 2]. In [Game 2], [the rule of the game]." The experimenter repeated this when switching from Game 2 to Game 3. As was the case

in Bock et al. (2015), the first three games were played to teach the rules to children and were not analyzed.

After three single-rule games, children were told to play one last game, which was harder to play. The final game is the border game. Border game is more difficult since children must switch between the previously learned single rules depending on the cards' borders to succeed. As Bock et al. (2015) used, children in the current study saw cards without borders, with dashed borders, and with full borders. The experimenter introduced the cards and rules to the children. Following a demonstration for each border type with verbal explanations, the experimenter reminded the rule one last time. There were twelve cards in Game 4, the border game. Children were scored 1 for each correctly sorted card, to a maximum of 12. See Appendix F for sample cards.

CHAPTER III

RESULTS

Three 7-year-olds could not pass the manipulation check for Condition 1 of the calibration task and were excluded from all analyses. The final sample consisted of 91 children. No outliers were detected in either the calibration or DCCSm tasks. All participants were within a three-SD range for all tasks.

3.1. Analysis Plan and Scoring

For all analyses, children's Endorsement score was treated as a categorical variable. The Trustworthiness score was measured as a continuous variable, and children's performance on Hypothetical Thinking was measured as a categorical variable. A total score on Hypothetical Thinking questions was also created. This total score was a continuous variable and ranged from 0 to 4. The order of the conditions was added to the analyses. Order has two levels: children saw Condition 1 first or Condition 2 first.

The Results section has four main parts. First, preliminary analyses are presented. The following three sections include analyses for Endorsement, Trustworthiness, and Hypothetical Thinking questions separately. Lastly, children's performance on the explicit and the implicit tasks are compared.

3.2. Preliminary Analyses

No significant effect of Gender and SES were found for calibration tasks, and they were not included in the following analyses. Table 2, Table 3, and Table 4 show bivariate correlations between Age in months, DCCSm, and the three calibration measures.

Overall, age was not significantly correlated with children's calibration. Age was only significantly related to children's Trustworthiness for the poorly calibrated accurate informant, such that older children found her less trustworthy (r = -.26, p = .02). Children's performance on DCCSm was not generally related to the dependent measures. DCCSm was only positively significantly related to children's success in Hypothetical Thinking for the well-calibrated hesitant informant in Condition 2—the condition where there is no accurate informant (r = .25, p = .01).

Children's Endorsement for Conditions 1 and 3 (i.e., conditions with a poorly calibrated and a well-calibrated accurate informant, respectively) were positively correlated (r = .33, p < .001). Children's Trustworthiness toward the well-calibrated hesitant informants across all conditions was highly correlated (for Condition 1 and

Table 2. Bivariate Correlations for Endorsement

Variables	1	2	3	4
1. Age (months)				
2. DCCSm	.14			
3. Condition 1 Endorse	-.04	.05		
4. Condition 2 Endorse	-.06	.17	.11	
5. Condition 3 Endorse	.08	.07	.33***	.18+

* $p < .05$, ** $p < .01$, *** $p < .001$, + $p < .01$

Table 3. Bivariate Correlations for Trustworthiness

Variables	1	2	3	4	5	6	7
1. Age (months)							
2. DCCSm	.14						
3. Condition 1 Well-Calibrated Hesitant Trust	-.20+	.10					
4. Condition 2 Well-Calibrated Hesitant Trust	-.04	.19+	.47***				
5. Condition 3 Well-Calibrated Hesitant Trust	-.01	.11	.60***	.60***			
6. Condition 1 Poorly Calibrated Accurate Trust	-.26*	-.06	-.06	-.09	-.05		
7. Condition 2 Poorly Calibrated Confident Trust	.082	-.01	.19+	.22*	.18+	-.26*	
8. Condition 3 Well-Calibrated Confident Trust	-.13	-.19+	.11	-.24*	-.38***	.12	.16

* $p < .05$, ** $p < .01$, *** $p < .001$, + $p < .01$

Table 4. Bivariate Correlations for Hypothetical Thinking

Variables	1	2	3	4	5
1. Age (months)					
2. DCCSm	.14				
3. Condition 1 Hesitantly Well-Calibrated HT	-.04	-.02			
4. Condition 2 Hesitantly Well-Calibrated HT	.14	.25*	.08		
5. Condition 3 Hesitantly Well-Calibrated HT	.03	-.15	.43***	.13	
6. Condition 3 Confidently Well-Calibrated HT	.12	.01	.19+	.11	.46***

* $p < .05$, ** $p < .01$, *** $p < .001$, + $p < .01$

Condition 2, r = .47; for Condition 1 and Condition 3, r = .60; for Condition 2 and Condition 3, r = .6, ps < .001).

3.3. Children's Endorsement Preference

Endorsement preferences were analyzed using binary logistic regressions. Children received a score of 1 if they endorsed the accurate and confident informant, and they received a score of 0 if they endorsed the well-calibrated hesitant informants or did not endorse any informant. Three binary logistic regressions were run for the Endorsement question (one for each condition). For each model, Endorsement was entered as the dependent variable. Age (7-, 9-, 11-year-old) and Order (Condition 1 first, Condition 2 first) were entered as predictors. These analyses were used to answer two research questions. First, whether older children are better at calibration than younger children in the school-age years (Research Question 1). Also, whether accuracy or confidence has a primary role in calibration (Research Question 2).

For Condition 1, the overall model was not significant (χ^2 (3, N = 91) = 1.54, p > .05). Children's Age or the Order in which the pairs were presented did not predict their Endorsement preferences in Condition 1. For Condition 2, the overall model was significant (χ^2 (3, N = 91) = 8.25, p < .05). The overall model explained 11.6% of the variance (Nagelkerke R Squared). Age was not a significant predictor of children's Endorsement decisions, but Order was. The odds ratio for Order was 2.73, which suggests that children were almost three times more likely to mistrust the poorly calibrated informant in the Endorsement question if Condition 2 was presented second (95% CI [1.14 – 6.53]).

For Condition 3, the overall model was significant (χ^2 (3, N = 91) = 9.84, p = .020). The model explained 18.3% of the variance. Age was not a significant predictor in the model, but Order predicted children's endorsement in Condition 3 was. The odds ratio for Order was .155 (95% CI [0.03 – 0.76]). This meant that children were more than six times more likely to trust the well-calibrated confident informant if children saw Condition 2 first and Condition 1 second.

A Cochran's Q test was run for all Endorsement questions to see whether children scored similarly for each Endorsement question or whether their responses were proportionally different across conditions. Cochran's Q revealed a significant difference in children's Endorsement across the three conditions ($\chi^2 = 42.43$, $p < .001$). Follow-up McNemar's tests were conducted for each pair. An exact McNemar's test between Condition 1 and Condition 2 Endorsement revealed that children avoided the poorly calibrated confident informant (IC) twice as much as the poorly calibrated accurate informant (AH; $p = .044$). The exact McNemar's test for Condition 1 and Condition 3 revealed that children endorsed the well-calibrated confident informant significantly more than the poorly calibrated accurate informant ($p < .001$). Another exact McNemar's test for Condition 2 and Condition 3 showed the same trend. Children endorsed the well-calibrated confident informants significantly more than they avoided endorsing the poorly calibrated confident informants ($p < .001$). See Table 5 for the descriptive statistics. Because all pairs were found to be significantly different, Endorsement scores were not collapsed for a total score.

3.4. Children's Trustworthiness Judgment

To examine children's judgments of Trustworthiness, two sets of Mixed ANOVAs were conducted. In the Condition set, two Mixed ANOVAs were conducted. One analysis was run with the well-calibrated confident informants (IH). The second ANOVA included the other three informants. In the Informant set, three Mixed ANVOAs were conducted. Informant pairs in each condition were paired in these ANOVAs. Similar to Endorsement, these analyses answered questions about age and the role of cues in calibration (Research Questions 1 and 2).

3.4.1. Trustworthiness Judgment for Conditions

Two 2 x 3 x 3 Mixed ANOVAs for Trustworthiness judgments were conducted for well-calibrated hesitant informants (IH), the within-subject factor was Condition (3 levels: Condition 1, Condition 2, Condition 3), and between-subjects factors were Age (3 levels: 7, 9, 11) and Order (2 levels: Condition 1 first, Condition 2 first). No main effect or interaction was significant for the well-calibrated hesitant informants (*ps* > .05).

To analyze children's Trustworthiness judgments for the other informants in each Condition, the within-subject factor was Condition (3 levels: Condition 1 (AH), Condition 2 (IC), Condition 3 (AC)), and between-subjects factors were Age and Order. There was a significant effect of Condition (F (2, 84) = 93.30, $p < .001$, η_p^2 = .69), and interaction between Condition and Order (F (1, 85) = 14.46, $p < .001$, η_p^2 = .26). No main effect of age was found. Pairwise comparisons for Condition revealed that children rated the poorly calibrated confident informant in Condition 2 (IC) less

trustworthy than the poorly calibrated accurate informant in Condition 1 (AH) (mean difference = 2.04, SE = .19, p < .001), and the well-calibrated confident informant in Condition 3 (AC) (mean difference = 2.35, SE = .17, p < .001). Children also rated the well-calibrated accurate informants (AC) more trustworthy than the poorly calibrated accurate informant (AH; mean difference = .33, SE = .12, p = .009).

Pairwise Comparisons for Pair and Order interaction showed that when children saw Condition 1 first (where the poorly calibrated accurate informant was in), they judged the poorly calibrated confident informant (IC) significantly less trustworthy (mean difference = 1.69, SE = .31, $p < .001$). There was also a marginally significant effect of Order when children saw Condition 2 first (where the confidently poorly calibrated informant, IC, was in), such that children found the poorly calibrated accurate informant (AH) more trustworthy (mean difference = .32, SE = .17, $p = .062$).

3.4.2. Trustworthiness Judgment for Informants

There were three 2 x 2 x 3 Mixed ANOVAs in the second set of Trustworthiness judgment analysis. For each informant pair, the Trustworthiness score of both informants was entered as a within-subject factor. Age and Order were between-subject factors. Analysis for Condition 1 showed that there is a main effect of informant (F (1, 85) = 175.41, $p < .001$, $\eta_p^2 = .67$), such that children found the poorly calibrated accurate informant (AH) more trustworthy than the well-calibrated hesitant informant (IH). There was a significant effect of Age (F (2, 85) = 7.54, $p = .006$, $\eta_p^2 = .11$) and Order (F (1,85) = 6.45, p = .035, $\eta_p^2 = .05$). The Order effect

showed that children scored informants in Condition 1 as more trustworthy when they saw Condition 1 first (Figure 3).

For Age, pairwise comparisons were conducted. Pairwise comparisons revealed that, on average, 7-year-old children rated informants as more trustworthy than 9-year-old children (mean difference = .62, SE = .22, p = .006) and 11-year-old children (mean difference = .61, SE = .21, p = .005). Specifically, 7-year-old children rated the well-calibrated hesitant informant (IH) significantly more trustworthy than 9-year-old children (mean difference = 1.05, SE = .40, p = .011). On the other hand, 7-year-old children rated the poorly calibrated accurate informant (AH) more trustworthy than 11-year-olds did (mean difference = .55, SE = .21, p = .010). There was no significant difference between children's Trustworthiness scores of the older age groups ($p > .05$).

Analysis for Condition 2 revealed a main effect of Informant (F (1, 85) = 4.59, p < .035, η_p^2 = .05) and an interaction between Informant and Order (F (1, 85) = 13.81, p < .001, η_p^2 = .14). Pairwise comparisons showed that children's trustworthiness judgments were not significantly different for the well-calibrated hesitant informant (IH) in Condition 2 ($p > .05$). Conversely, there was a significant difference between children's trustworthiness judgments for the poorly calibrated confident informant (IC), such that children said this informant is more trustworthy when Condition 2 was presented first (mean difference = 1.69, SE = .31, $p < .001$; see Figure 2).

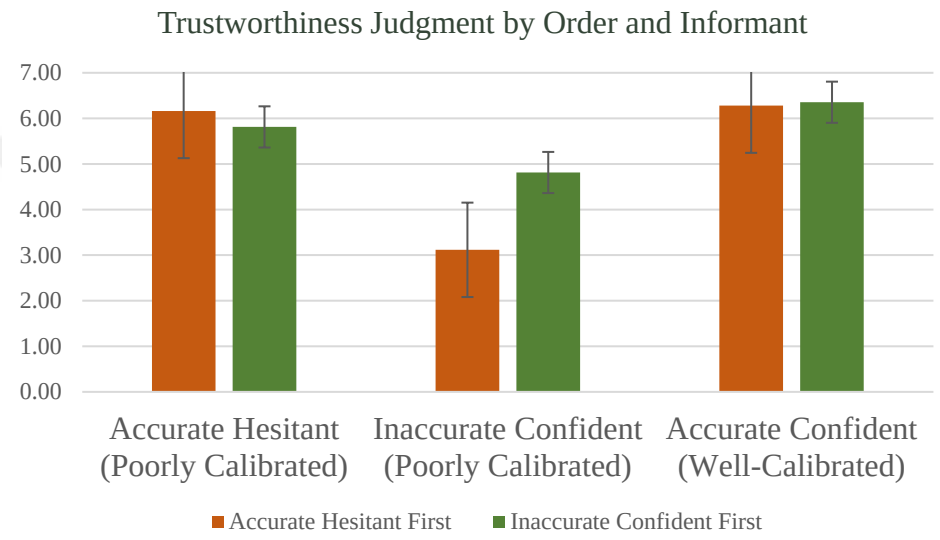
Table 5. Descriptive Statistics of Children's Endorsement and Their Success in Endorsement

	Condition 1				Condition 2				Condition 3			
Order	IH	Neither	AH	Pass	IH	Neither	IC	Pass	IH	Neither	AC	Pass
Condition 1 First	13 (30.2%)	1 (2.3%)	29 (67.4%)	14 (32.5%)	10 (23.3%)	7 (16.3%)	**26 (60.5%)**	**17 (39.6%)**	2 (4.7%)	0 (0.0%)	41 (95.3%)	41 (95.3%)
Condition 2 First	14 (29.2%)	5 (10.4%)	29 (60.4%)	19 (39.6%)	16 (33.3%)	14 (29.2%)	**18 (37.5%)**	**30 (62.5%)**	7 (14.9%)	3 (6.4%)	37 (78.7%)	37 (78.7%)
Total	27 (29.7%)	6 (6.6%)	58 (63.7%)	33 (36.3%)	26 (28.6%)	21 (23.1%)	44 (48.4%)	47 (51.6%)	9 (10.0%)	3 (3.3%)	78 (85.7%)	78 (85.7%)

Note. Bolded values indicate significant proportion differences.

Analysis for Condition 3 showed that there was a main effect of Informant (F (1, 85) = 155.53, $p < .001$, $\eta_p^2 = .65$), such that children rated the well-calibrated confident informant (AC) more trustworthy than the well-calibrated hesitant informant (IH). There was no other main effect or interaction ($ps > .05$).

Figure 3. Children's Trustworthiness Judgments



Note. Children's Trustworthiness for informants in Condition 1 (left), Condition 2 (middle), and Condition 3 (right).

3.5. Children's Performance on Hypothetical Thinking Questions

Children passed or failed each Hypothetical Thinking question. To investigate age effects, four separate binary logistic regressions were conducted to observe children's performance on each Hypothetical Thinking question by age. To detect any order effect, regression analyses were run in two blocks. The first block included Order only. The second block included Order and Age. As stated above, only well-calibrated informants are included in this analysis (i.e., inaccurate and hesitant

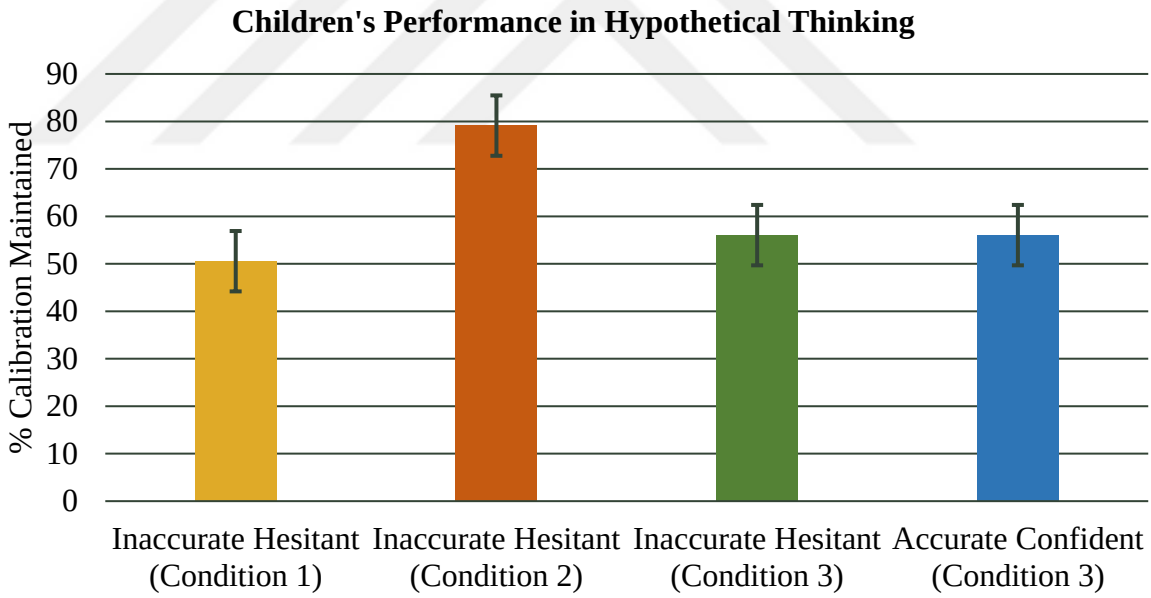
informants (IH) in Condition 1, Condition 2, and Condition 3 and for the accurate and confident informant (AC) in Condition 3).

The model for Hypothetical Thinking for IH in Condition 1 showed that Order and Age were not significant predictors of children's performance (χ^2 = 2.20, p = .33). However, the model was significant when Age was removed (χ^2 = 4.09, p = .043). The odds ratio for Order was .42. This indicated that children were approximately 2.4 times more likely to succeed on Hypothetical Thinking for IH in Condition 1 if children first saw Condition 2. That is, there is an order effect on children's Hypothetical Thinking in Condition 1. The model for IH in Condition 2, Age, and Order were not significant (χ^2 = 2.70, p = .26). Children's Age and the Order in which they were presented with the informant pairs did not predict their performance on Hypothetical Thinking for IH in Condition 3 (χ^2 = .89, p = .64), and for AC in Condition 3 (χ^2 = 1.71, p = .43).

A Cochran's Q test was conducted to investigate whether four questions are meaningfully different from each other. The test determined that children's performance for separate Hypothetical Thinking proportions was significantly different (χ^2 = 24.12, p < .001). As a follow-up, McNemar enabled a two-by-two comparison for which of the pairs were significantly different. McNemar tests revealed that children's performance on Hypothetical Thinking was nott proportionally different from each other among all pairs among the well-calibrated hesitant informant (IH) in Condition 1, Condition 3, and the well-calibrated confident informant AC in Condition 3 (*ps* > .05). McNemar tests also showed that children's

performance on Hypothetical Thinking questions for the well-calibrated hesitant informant (IH) in Condition 2 was proportionally different from IH in Condition 1 (*asymptotic p* < .001), IH in Condition 3 (*asymptotic p* < .001), and AC in Condition 3 (*asymptotic p* = .003). Accordingly, children succeed in Hypothetical Thinking more for the well-calibrated hesitant informant (IH) in Condition 2 than the other three informants (Figure 4). This indicates that children perform better in the Hypothetical Thinking question for Condition 2, and the absence of accuracy evokes sensitivity toward calibration.

Figure 4. Performance on Hypothetical Thinking Questions for Well-calibrated Informants



Although four of the Hypothetical Thinking questions were not correlated, a total score for Hypothetical Thinking was generated. A Two-Way ANOVA with Age and Order entered as factors showed that the effect of age was not significant (F (2, 85) =

1.12, p = .33). The effect of order (F (1, 85) = .91, p = .34). and interaction between Age and Order (F (2, 85) = .75, p = .47) were not significant.

3.6. Implicit and Explicit Calibration

Hypothetical Thinking and Trustworthiness were conceptualized as implicit and explicit calibration understanding, respectively. We used McNemar's tests for Conditions 1 and 2 to test whether Hypothetical Thinking precedes Trustworthiness. Children's Hypothetical Thinking pass or fail score was used. For Trustworthiness, children were considered passing if they rated the poorly calibrated informants less trustworthy than the well-calibrated hesitant informants. The results showed that there was a significant difference in the proportions of children who passed one task but failed the other, such that children passed Hypothetical thinking more (ps < .001).

3.7. Summary of Results

For Endorsement, no significant effect of Age or Order was found. In Condition 2, children endorsed the poorly calibrated confident informant (IC) less when Condition 2 was presented second. In Condition 3, children endorsed the well-calibrated confident informants when they saw Condition 2 first. A comparison between Endorsement questions for each condition revealed no equivalence between children's preferences, such that children preferred the well-calibrated confident informant (AC) more than they avoided endorsing the poorly calibrated informants in Conditions 1 and 2. Also, children avoided endorsing the poorly calibrated confident

informant in Condition 2 (IC) more than the poorly calibrated accurate informant in Condition 1 (AH).

Regarding Trustworthiness judgments, there was no effect of Condition, Age, or Order for well-calibrated hesitant informants (IHs) in all conditions. Condition analyses for the other set of informants showed that children rated the poorly calibrated confident informant (IC) as less trustworthy than the poorly calibrated accurate informant (AH) and the well-calibrated confident informant (AC). Additionally, children found the poorly calibrated confident informant (IC) less trustworthy when children saw Condition 1 first, which included the poorly calibrated accurate informant (AH). There was also a marginal difference in children's Trustworthiness judgment for the poorly calibrated accurate informant (AH) when children saw Condition 2 first, which included the poorly calibrated confident informant (IC), such that children rated the poorly calibrated accurate informant (AH) as more trustworthy.

Informant analyses for Trustworthiness judgments showed that children rated the well-calibrated hesitant informants (IHs) as less trustworthy than any of the informants that they were paired within each condition. Seven-year-old children had higher Trustworthiness judgment than 9- and 11-year-olds in Condition 1. More specifically, 7-year-old children rated the well-calibrated hesitant informant (IH) more trustworthy than 9-year-olds, and they rated the poorly calibrated accurate informant (AH) more trustworthy than 11-year-olds. For Condition 2, children judged the poorly calibrated confident informant (IC) more trustworthy when

Condition 2 was first. For Condition 3, no difference other than Informant was found.

Analyses for Hypothetical Thinking showed that children passed the Hypothetical question for Condition 1 more when Condition 2 was first (before controlling for Age). Age and Order did not predict children's performance on Hypothetical questions for Conditions 2 and 3. Children's responses to three of the four Hypothetical Thinking questions were similar. Children succeeded in Hypothetical Thinking for Condition 2 significantly more than Conditions 1 and 3. A combined score of Hypothetical Thinking (range 0-4) showed no effect of Age or Order.

The explicit Trustworthiness task was more challenging for children, and they were better at the implicit Hypothetical Thinking task. Lastly, children's EF, as measured by DCCSm, was not related to their performance on the calibration tasks. The exception to the null results was the positive correlation between DCCSm and children's score on Hypothetical Thinking for Condition 2.

CHAPTER IV

DISCUSSION

4.1. Research Questions

4.1.1. Age Effects (Research Question 1)

The aim of the current study was to investigate children's understanding of calibration and how it is assessed. To that end, the current study introduced new methodologies appropriate for the sample age. Accordingly, the current study is partly exploratory. The current study extended the age at which researchers have looked at children's understanding of hesitance calibration. Contrary to expectation, children's calibration understanding did not show a robust improvement by age from 7 to 11 years. No robust age effect was found for any of the tasks.

The lack of improvement by age has two possible explanations. First, understanding hesitance calibration may be a later developing ability than hypothesized in the current study. While this may be true, it begs the question of explaining differences

in calibration measures because the results for Endorsement, Trustworthiness, and Hypothetical Thinking were not uniform. The second explanation acknowledges the multifaceted nature of calibration and its operationalization. Whether people calibrate also depends on other factors, such as cognitive load (Tenney et al., 2011). The multifaceted nature of calibration is further elaborated below in relation to other research questions. Prior to that, one note should be added.

In the current study, the age of 9 was assumed as a turning point in children's calibration understanding. Children's understanding of "virtuous ignorance" motivated this expectation (i.e., the understanding that it is okay for experts in a field to *not* know information that is hard to know). Previous research showed that 5-year-olds could differentiate hard-to-know questions from easier ones, but it was not until the age of 9 that children thought it was okay for experts not to know hard questions (Kominsky et al., 2016). The distinction between being accurate (or confident) and well-calibrated and knowing easy-to-know information and not knowing hard-to-know information was considered equivalent. Calibration and virtuous ignorance require children to think in terms of double-negation about an epistemic situation: in calibration, children should recognize that it is okay to be hesitant when inaccurate because this is what someone who is merely guessing would do; in virtuous ignorance, children should recognize that it is okay not to know something, even if you are an expert in the area.

However, there is one difference between calibration and virtuous ignorance. In calibration, the informant does not say anything directly about her own credibility

(i.e., calibration). Instead, the listener has to infer that the informant is well-calibrated or poorly calibrated. In contrast, for virtuous ignorance, children had already stated the difficulty of the questions. That is, recognition of calibration is indirect, but virtuous ignorance is direct. How complex or salient calibration and virtuous ignorance cues may be influenced by their directness, therefore, may have different developmental milestones. Since indirect cues may be harder to detect, it is understandable that a robust understanding of calibration comes later.

4.1.2. Accuracy versus Confidence (Research Question 2)

A second hypothesis predicted an asymmetry between accuracy and confidence in children's calibration understanding, such that one cue might be more salient than the other. As expected, a comparison of poorly calibrated informants revealed that children endorse the poorly calibrated accurate informant more than the poorly calibrated confident informant. Moreover, children's lack of discrimination between poorly calibrated and well-calibrated accurate informants (i.e., no confound between calibration and accuracy) in Trustworthiness suggests that children do not show calibration understanding[5]. These findings show that accuracy is not only a more salient cue, but it can also trump calibration.

Findings for Hypothetical Thinking parallel the results for Trustworthiness but from the opposite direction. In Hypothetical Thinking, children performed at chance level for informants in Conditions 1 and 3. However, children were significantly higher than the chance level in Condition 2 in their Hypothetical Thinking. This finding

[5] Children differentiated accurate informants in Endorsement but not in Trustworthiness.

showed that thinking about calibration may not be evoked in the presence of accuracy. Instead, only in the absence of an accurate informant (as in Condition 2) do children start to take calibration into account. In short, the findings from the Hypothetical Thinking measure led to the conclusion that *inaccuracy evokes calibration*, while *accuracy trumps calibration*. These conclusions are compared in light of the other results below.

Although we did not expect an effect of Order, the results further inform us about the strength of accuracy when task Order varies. In the Endorsement task, children avoided the poorly calibrated confident informant in Condition 2 more if they saw Condition 1 first. This shows that children were more likely to avoid endorsing the poorly calibrated confident informant if they saw an accurate informant beforehand. In the Trustworthiness task, children rated the poorly calibrated confident informant in Condition 2 less trustworthy if they saw Condition 1 first. Similarly, this shows that children were more influenced by seeing an accurate informant. Hence, seeing an accurate informant has a strong effect on children's future judgments about the inaccurate informants.

Overall, the three dependent measures for calibration showed that children between the ages of 7 and 11 show similar degrees of sensitivity to calibration. In terms of the measures, there were some differences. For Endorsement, children did not systematically avoid endorsing the poorly calibrated informants. They only differentiated the well-calibrated and poorly calibrated accurate informants, though this was not the case for Trustworthiness. In Trustworthiness, only the poorly

calibrated confident informant was rated less trustworthy. In Hypothetical Thinking, where an effect of condition was not expected, indicators of children's understanding of calibration were found for Condition 2 (i.e., the condition in which there was a poorly calibrated confident informant). Also, Order effects on the Endorsement and Trustworthiness tasks suggest that children are affected by having seen an accurate informant before its confident counterpart, but not vice versa. These results lead to the conclusion that accuracy is a stronger cue in children's calibration understanding, although children do not have a complete understanding of calibration at this age.

4.1.3. Implicit and Explicit Understanding of Calibration (Research Question 3)

A third hypothesis compared Trustworthiness and Hypothetical Thinking tasks as measures of explicit and implicit calibration understanding, respectively. Results showed that children were more challenged by the Trustworthiness questions than the Hypothetical Thinking questions for Conditions 1 and 2. Although the results are clear, one should be cautious about the interpretation. Trustworthiness was considered an explicit measure of calibration because children directly answered how credible they thought the informants were. Hypothetical Thinking was considered an implicit measure of calibration because children had to think in terms of possible double negation operations. So, children who showed an understanding of calibration had to perform an explicit double negation operation, which in turn corresponded to an implicit understanding of calibration. So perhaps Hypothetical Thinking involves both implicit and explicit aspects.

4.1.4. Calibration and EF (Research Question 4)

The current study was motivated by the idea that calibration is a complex phenomenon. The complexity of calibration lies in the intertwined nature of epistemic and social factors and the double negation required in hesitance calibration to infer that someone hesitant and inaccurate is still well-calibrated. EF, specifically working memory, was considered a relevant factor that could affect children's performance on calibration understanding. Contrary to the expectation, children's performance on the calibration measures was not consistently related to EF. Only DCCSm and children's performance on the Hypothetical Thinking task in Condition 2 were positively correlated. However, these findings should be approached cautiously for three reasons.

First, Age and children's DCCSm scores were not correlated. One possible reason for the sample discrepancy may be task difficulty. The original DCCS is a standard cognitive flexibility/switching task for preschoolers. As such, DCCS signifies a competence factor of flexibility (i.e., DCCS is a measure of cognitive flexibility, which show whether children's capability to perform the task in general). DCCSm is an adaptation of DCCS, but the working memory load of DCCSm might trump the competence factors of DCCS (i.e., DCCSm includes cognitive flexibility, but working memory ability is also determinant). Working memory could have been a stronger performance factor than anticipated, such that it is unclear whether children can succeed with this particular version of the task. If this is the case, children in the current study did not show an improvement by age because DCCSm played the role of a performance factor (i.e., working memory) that was too challenging for them.

The second reason to be cautious about EF in the current study is that EF results in the current study do not align with Bock and colleagues' (2015) findings. In that study, it was found that children performed better, and their performance increased with age. Neither was the case in the current study. Unlike the original DCCS, DCCSm is not a commonly established and consensually validated task. To the best of our knowledge, there is no other study that utilizes Bock and colleagues' version of DCCS. Also, the DCCS adaptation in the current study was not a direct translation requested from the authors. Instead, the modified version was prepared by modifying the original version of DCCS in light of Bock and colleagues' descriptions. Thus, there might have been slight differences in the two applications of the task such that administration was less detailed in the current study. This could have led to a variation in the task affecting children's performance.

The third reason concerns sampling. Bock and colleagues (2015) had an American sample[6]. The difference in DCCSm performance may also be due to a cultural difference. The current study and Bock and colleagues cannot address that issue. However, there is indirect evidence that suggests there is no significant difference in children's EF abilities in Turkish and Western populations. Lucas et al. (2013) showed that British and Turkish preschoolers were not different from each other in their EF performance, although Chinese children performed higher than both groups. Though the present evidence cannot rule out this alternative, it is unlikely that the cultural variability explains the difference.

[6] The authors do not refer to their sample directly. It is inferred from their participants section.

Cognitive flexibility as a component of EF was considered a competence factor because calibration (Hypothetical Thinking specifically) requires attention and a shift between multiple aspects of the same subject. On the other hand, working memory was considered a performance factor for calibration because calibration requires inferences of its constituent cues. Bock and colleagues' (2015) modified version of DCCS was one plausible task to use in the current study because it is a cognitive flexibility task with more working memory demand than the original DCCS. Null results for EF suggest that either the logic does not generally hold for EF or a battery of EF tasks would better inform any possible relations to calibration understanding. At this point, the latter option seems more appropriate to pursue. To better examine the relevance of EF to calibration, future research should consider using more EF measures.

4.2. Endorsement

Endorsement questions are well-established in selective trust research and are widely used with preschoolers (Tong et al., 2020). The current study was conducted with school-aged children to assess better hesitance calibration. The endorsement question used in the current study had one difference from Birch and colleagues' (2020) endorsement question. Children could choose not to endorse either informant in the current study. This was done because informants in conditions were not always preferable, like the well-calibrated confident informant. For Conditions 1 and 2, being sensitive to calibration meant children avoided endorsing the poorly calibrated informants. The third option lets children not endorse the poorly calibrated informant while also not endorsing the other informant.

It should be remembered that studying confidence calibration has potential confounds (Figure 1). Showing a preference toward the well-calibrated informant does not necessarily imply calibration understanding (i.e., it could just be about accuracy). Given the challenges of studying calibration with a selective trust paradigm, researchers should consider using the original witness paradigm adapted by Tenney and colleagues (2011) or developing new calibration methods.

4.3. Limitations and Future Directions

Despite the methodological novelties in the current study, there are several limitations. First, although the current study helped advance calibration research, it turns out not to have been optimally sampled. The oldest age group, 11-year-olds, was slightly younger than what may be ideal.

The current study did not include an adult group. Since each dependent measure was new or modified, a baseline group including adults would enable a better interpretation of the results. Literature shows that adults have calibration understanding, but not always. An adult sample would enable a direct comparison between children and adults.

The current study had methodological changes but did not directly compare standard and modified versions of the tasks. Although the methodological changes in the current study were justified, future research should consider including old and new

measures simultaneously to determine whether the changes in the current study were as meaningful as argued. Specifically, the sensitivity to Endorsement in relation to calibration may be a later developing ability, or Endorsement may not be the ideal measure for understanding hesitance calibration for this age group. The most viable possibility for why Endorsement did not work as expected is the asymmetry between confidence and hesitance calibration (i.e., in one, accuracy is held constant, in the other, it varies). That is, Endorsement may be a good measure for well-calibrated confident informants but not for well-calibrated hesitant informants. Future research can account for these changes.

As stated above, the EF measure in the current study was a little-used modification of DCCS. Given the null results, the construct validity of DCCSm is questionable. A step forward is to use more well-established measures of EF for school-aged children. Of specific interest, cognitive flexibility and working memory should be separately assessed to capture both EF components directly.

Because understanding hesitance calibration requires double-negation, another relevant measure to control for is double-negation. To the best of our knowledge, there is no standardized task for double negation in child development. Double-negation tasks are usually linguistic and do not target cognitive development (e.g., Jou, 1988). If devised, a double-negation measure for children's general cognitive development would be informative to investigate in addition to EF.

A third cognitive ability that may be relevant to calibration is Theory of Mind (ToM). The current study could not include EF and ToM due to time constraints. ToM, in addition to EF, still remains a potentially relevant cognitive ability for calibration.

To support the relevance of ToM, previous research suggest that selective trust and false belief understanding are related in the preschool age (DiYanni et al., 2012; Lucas et al., 2013). However, ToM was not found to be wholly relevant to selective trust (Allen et al., 2021). Since calibration is considered a sub-area of selective trust, ToM may be more relevant than EF. In fact, Lucas and colleagues (2013) demonstrated that ToM but not EF was related to flexible trust. In short, ToM may be related to calibration for the same reason selective trust is assumed to be related to ToM. However, ToM development in the school-age period is not uniform (Osterhaus et al., 2016). One difficulty in including ToM to calibration research is to decide which of the aspects of advanced ToM to assess (e.g., Higher-order False-belief tasks, Strange Stories, interpretive ToM, etc.).

4.4. To Explain Calibration

The current study contributed to the empirical study of calibration. Some conceptual decisions were made to that end, such as focusing on hesitance calibration. However, the current study did not address a framework to explain calibration. An account of calibration should clarify if accuracy trumps calibration and/or if inaccuracy evokes calibration. If both can be the case, it should address which one is applicable, when, and why.

First, understanding of calibration has relevant competence and performance components. Whether someone *can* calibrate is competence-relevant and is signified by development (i.e., it is about whether someone can calibrate at all). Whether someone *does* calibrate is performance-relevant and varies depending on situational constraints. To exemplify, Tenney and colleagues (2011) showed 5- and 6-year-old children did not calibrate because they cannot calibrate. They also showed that adults do not calibrate when under cognitive load because that specific situation makes them unable to calibrate.

In these examples, competence factors converge with *inaccuracy evokes calibration*, and performance factors converge with *accuracy trumps calibration*. The former suggests that accuracy fosters the grasp of calibration in a situation. The latter suggests that calibration can be overrun with accuracy alone. For calibration to be overrun, it must be present. That is, one should be competent in calibration for it to be trumped. On the other hand, inaccuracy evoking calibration is like a scaffold for people to be sensitive to calibration before it is fully developed. Such sensitivity would be trivial if people were already able to calibrate. While a complete framework for calibration goes beyond the scope of the current study, these implications should be accounted for in an explanatory framework for calibration understanding.

4.5. Conclusion

The main aim of the current study was to investigate children's understanding of others' trustworthiness. To that end, the current study expanded research on

calibration with a wider age range and a novel set of calibration measures. The results showed that children displayed some levels of calibration understanding in different situations, but there was no age differences. While accuracy emerged as the cue that has more weight for calibration, whether accuracy trumps calibration or inaccuracy evokes calibration (or both) remains to be more fully answered. Also, the current study concluded—with caution—that children performed better in the implicit measure of calibration (Hypothetical Thinking). Regarding EF, performance was not related to calibration for the most part. Finally, the current study pushed calibration research one step forward. It filled some gaps but opened doors to new questions. That is why the conclusions from the current study have been more excogitative than conclusive. Calibration understanding has important implications for children since fake news sometimes spreads faster than actual news. Calibration can help increase vigilance toward fake news and misinformation of other kinds.

REFERENCES

Allen, J. W., Sümer, C., & Ilgaz, H. (2021). Cultural affordances: Does model reliability affect over-imitation in preschoolers. *Cognitive Development*, *57*, 100999. https://doi.org/10.1016/j.cogdev.2020.100999

Audi, R. (1997). The place of testimony in the fabric of knowledge and justification. *American Philosophical Quarterly, 34*(4), 405-422.

Birch, S. A., Severson, R. L., & Baimel, A. (2020). Children's understanding of when a person's confidence and hesitancy is a cue to their credibility. *PloS One*, *15*(1), e0227026. https://doi.org/10.1371/journal.pone.0227026

Bock, A. M., Gallaway, K. C., & Hund, A. M. (2015). Specifying links between executive functioning and theory of mind during middle childhood: Cognitive flexibility predicts social understanding. *Journal of Cognition and Development, 16*(3), 509-521. https://doi.org/10.1080/15248372.2014.888350

Brosseau-Liard, P., Cassels, T., & Birch, S. (2014). You seem certain but you were wrong before: Developmental change in preschoolers' relative trust in accurate versus confident speakers. *PloS One*, *9*(9), e108308. https://doi.org/10.1371/journal.pone.0108308

DiYanni, C., Nini, D., Rheel, W., & Livelli, A. (2012). ‘I won't trust you if I think you're trying to deceive me': Relations between selective trust, theory of mind, and imitation in early childhood. *Journal of Cognition and Development, 13*(3), 354-371. https://doi.org/10.1080/15248372.2011.590462

Fusaro, M., & Harris, P. L. (2013). Dax gets the nod: Toddlers detect and use social cues to evaluate testimony. *Developmental Psychology, 49*(3), 514-522. https://doi.org/10.1037/a0030580

Heyes, C. (2012). Grist and mills: On the cultural origins of cultural learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1599), 2181-2191. https://doi.org/10.1098/rstb.2012.0120

Jou, J. (1988). The development of comprehension of double negation in Chinese children. *Journal of Experimental Child Psychology, 45*(3), 457-471. https://doi.org/10.1016/0022-0965(88)90042-2

Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science, 15*(10), 694-698. https://doi.org/10.1111/j.0956-7976.2004.00742.x

Koenig M. A., & McMyler, B. (2019). Testimonial Knowledge: Understanding the evidential, uncovering the interpersonal. In M. Fricker, P. J. Graham, D. Henderson, & N. J. L. L. Pedersen (Eds.), *The Routledge Handbook of Social Epistemology* (pp. 103-114). Routledge.

Kominsky, J. F., Langthorne, P., & Keil, F. C. (2016). The better part of not knowing: Virtuous ignorance. *Developmental Psychology, 52*(1), 31-45. https://doi.org/10.1037/dev0000065

Lucas, A. J., Lewis, C., Pala, F. C., Wong, K., & Berridge, D. (2013). Social-cognitive processes in preschoolers' selective trust: three cultures compared. *Developmental Psychology*, *49*(3), 579-590. https://doi.org/10.1037/a0029864

Sporer, S. L. (1982). A brief history of psychology of testimony. *Current Psychological Review*, *2*(3), 323-339.

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science*, *18*(1), 46-50. https://doi.org/10.1111/j.1467-9280.2007.01847.x

Tenney, E. R., Small, J. E., Kondrad, R. L., Jaswal, V. K., & Spellman, B. A. (2011). Accuracy, confidence, and calibration: how young children and adults assess credibility. *Developmental Psychology*, *47*(4), 1065-1077. https://psycnet.apa.org/doi/10.1037/a0023273

Tenney, E. R., Spellman, B. A., & MacCoun, R. J. (2008). The benefits of knowing what you know (and what you don't): How calibration affects credibility. *Journal of Experimental Social Psychology*, *44*(5), 1368-1375. https://doi.org/10.1016/j.jesp.2008.04.006

Tong, Y., Wang, F., & Danovitch, J. (2020). The role of epistemic and social characteristics in children's selective trust: Three meta-analyses. *Developmental Science*, *23*(2), e12895. https://doi.org/10.1111/desc.12895

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523-541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature Protocols*, *1*(1), 297-301. https://doi.org/10.1038/nprot.2006.46

APPENDICES

Appendix A

Snapshots from documentary videos used:

Appendix B

Questions for each documentary

Kanada Geyikleri (Elks):

Soru 1: Kanada Geyikleri havalar ısınınca ne yaparlar?

Kanada Geyikleri havalar ısınınca ormanı terk ederler.

Soru 2: Kanada Geyikleri nasıl renklidir?

Kanada Geyiklerinin baş kısmı koyu, gövdeleri açık renklidir.

Soru 3: Kanada Geyikleri diğer geyiklerden büyük müdür yoksa küçük müdür?

Kanada Geyikleri yeryüzünün en büyük geyiklerindendir.

Soru 4: Kanada Geyikleri neyle beslenir?

Kanada Geyikleri yaprak, ağaç kabukları ve dal yerler.

Gümüş Yaprak Maymunları (Silvered Leaf Monkey)

Soru 1: Gümüş yaprak maymunları nerede yiyecek ararlar?

Gümüş Yaprak Maymunları yiyeceklerini ormanın kıyıya yakın yerinde ararlar.

Soru 2: Gümüş yaprak maymunları ne ile beslenirler?

Gümüş yaprak maymunları meyve ve lifli bitkilerle beslenirler.

Soru 3: Gümüş yaprak maymunlarının dişleri nasıldır?

Gümüş Yaprak Maymunlarının dişleri keskin uçludur.

Soru 4: Gümüş yaprak maymunları vakitlerini nerede geçirirler?

Gümüş yaprak maymunları vakitlerini ağaç üzerinde geçirirler.

Tokmak Başlı Bataklık Balıkları (Mudskippers)

Soru 1: Tokmak Başlı Bataklık Balıkları zamanlarının çoğunu nerede geçirir?

Tokmak Başlı Bataklık Balıkları zamanlarının çoğunu karada geçirirler.

Soru 2: Tokmak Başlı Bataklık Balıkları nasıl hareket ederler?

Tokmak Başlı Bataklık Balıkları bacağa benzeyen yüzgeçleri ile yürürler.

Soru 3: Tokmak Başlı Bataklık Balıklarının nasıl gözleri vardır?

Tokmak Başlı Bataklık Balıklarının her yöne dönebilen, küresel gözleri vardır.

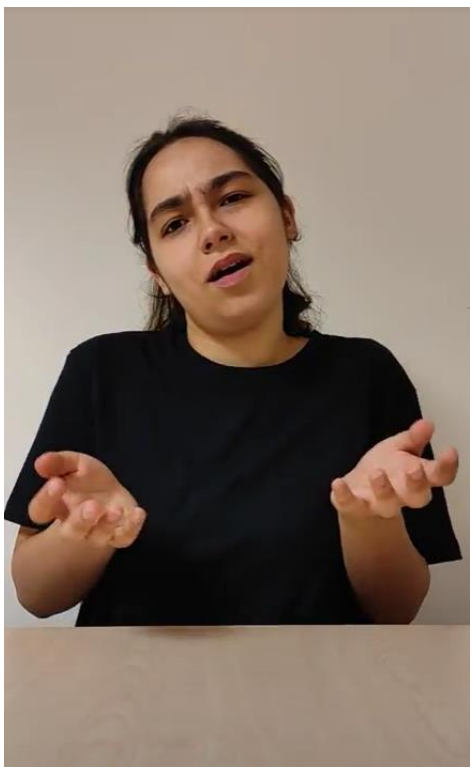
Soru 4: Tokmak Başlı Bataklık Balıkları derilerinin arasındaki keseciklerde ne depolarlar?

Tokmak Başlı Bataklık Balıkları derilerinin arasındaki keseciklerde su depolarlar.

Appendix C

Confident Informant Samples:

Hesitant Informant Samples:



Appendix D

“[Kaynağın ismi] bizim izlediğimiz videolarda kendinden emin olarak/tereddüt ederek konuşmuştu. Şimdi [Kaynak ismi]’nin [katılımcının en sevdiği hayvan] hakkında tereddüt ederek/emin olarak konuşacağını düşünelim. Sence [Kaynak ismi] bu defa doğru bir şey mi söyler, yanlış bir şey mi?”

Appendix E

The Trustworthiness scale shown to children, ranged from 1 (not trustworthy at all) to 7 (very trustworthy):

Appendix F

Card samples from DCCSm: two blue triangles (no border), four red circles (full border), six yellow squares (dashed border) in the order of appearance.