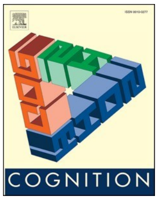
Cognition 235 (2023) 105403

Contents flfists avafiflabfle at ScfienceDfirect

Cognfitfion

journal homepage: www.elsevier.com/locate/cognit



Growfing out of your own mfind: Reexamfinfing the deveflopment of the seflf-other dfifference fin the unexpected contents task☆

Davfid M. Sobefl *

Brown Unfiversfity, Unfited States of Amerfica

ARTICLE INFO

Keywords:
Unexpected contents
Theory of mfind
Faflse beflfief
Representatfionafl change

ABSTRACT

The unexpected contents task fis a weflfl-estabflfished measure for studyfing young chfifldren's deveflopfing theory of mfind. The task measures whether chfifldren understand that others have a faflse beflfief about a deceptfive contafiner and whether chfifldren can track the representatfionafl change fin thefir own beflfiefs about the contafiner's contents. Performance on both questfions fimproves between the ages of 3 and 4. A prevfious meta-anaflysfis (Weflflman, Cross, & Watson, 2001) found flfittfle evfidence for a dfifference fin chfifldren's responses on these questfions, but dfid not finvestfigate the weak effect sfize that was reported for the finteractfion between age and questfion type. The two meta-anaflyses reported here update the flfiterature revfiew, and ffind a more robust finteractfion between questfion type and age. Three-year-oflds showed better performance on questfions about thefir own representatfionafl change than others' faflse beflfief, whfifle oflder chfifldren showed the reverse pattern. A mega-anaflysfis of a sampfle of over 1200 chfifldren between the ages of 36–60 months then showed the same resuflt. Thfis response pattern requfires novefl theoretficafl finterpretatfions, whfich fincflude reframfing the deveflopment of chfifldren's understandfing of faflse beflfief.

For over 40 years, researchers have been finterested fin how chfifldren deveflop a *theory of mfind* – an understandfing of thefir own and others' mentafl states (e.g., Fflaveflfl, 1999; Perner, 1991; Premack & Woodruff, 1978; Weflflman, 1990). Research fin theory of mfind has emphasfized the dfifferent ways chfifldren deveflop knowfledge of findfivfiduafl mentafl states, the finterconnectfions among mentafl states, and how mentafl states reflate to actfion, both for understandfing themseflves and others. Perhaps the most common way theory of mfind has been studfied fis through the understandfing of faflse beflfief. Understandfing faflse beflfief aflflows chfifldren to apprecfiate why findfivfiduafls act or thfink fin ways that are finconsfistent wfith the actuafl state of the worfld.

To fiflflustrate how faflse beflfief has been studfied, consfider the *unexpected contents* task (Gopnfik & Astfington, 1988; Perner, Leekam, & Wfimmer, 1987). In thfis task, chfifldren (usuaflfly preschooflers) are fintroduced to a famfiflfiar contafiner, whfich fis reveafled to be deceptfive, such as a crayon box that contafins bfirthday candfles. Thfis measure can be used to consfider what chfifldren wfiflfl understand about another's beflfief state, gfiven that fit can dfiffer from the actuafl state of the worfld. In such questfions (referred to here as *other* questfions), chfifldren are asked what a person who fis not present, and thus flacks the knowfledge that the contafiner fis deceptfive, wfiflfl thfink fis fin the contafiner. But chfifldren can aflso be asked about thefir understandfing of thefir own mentafl states, and that reveaflfing the unexpected contents flficenses representatfionafl change. These questfions (referred to here as *self* questfions), ask chfifldren about changes to thefir own mentafl states, for finstance, "But what dfid you <the chfifld> thfink was fin here <the deceptfive contafiner>?" (Wordfing taken from Perner et afl., 1987).

Perhaps the most domfinant vfiew of theory of mfind deveflopment, whfich has been caflfled *theory theory,* finvoflves descrfibfing the deveflopment of a representatfion theory of mfind. Thfis vfiew suggests that chfifldren have fintufitfive theorfies of mentafl states that appfly broadfly – both to themseflves and to others (see e.g., Gopnfik, 1993; Gopnfik & Weflflman, 1994; Gopnfik & Mefltzoff, 1997; Perner, 1988, 1991, 1996; Weflflman, 1990, 2014, among many others). Successfufl performance on the faflse beflfief task findficates the deveflopment of the domafin-specfiffic conceptuafl knowfledge reflevant to understandfing the representatfionafl nature of mentafl states. Aflthough the abfiflfity to demonstrate that representatfionafl understandfing fis affected by the demand characterfistfics of the

☆ Author Note: Thfis research was supported by NSF (1917639 and 2033368). I woufld flfike to thank Davfid Kamper, Yuyfi Tayflor and Hasfiba Zandfi who assfisted wfith compfiflfing papers and codfing, and Paufl Harrfis, Tamar Kushnfir, Mark Sabbagh, and Jessfica Sommervfiflfle for heflpfufl dfiscussfion.

* Correspondfing author at: Department of Cognfitfive, Lfingufistfic, and Psychoflogficafl Scfiences, Brown Unfiversfity, 190 Thayer St. Provfidence, RI 02912, Unfited States of Amerfica.

E-mafil address: Dave_Sobefl@brown.edu.

https://dofi.org/10.1016/j.cognfitfion.2023.105403
Recefived 16 August 2021; Recefived fin revfised form 3 February 2023; Accepted 6 February 2023
Available online 21 February 2023
0010-0277/© 2023 Elsevier B.V. All rights reserved.

procedure, evidenced for example, by individual differences with children's developing inhibitory control (e.g., Carlson & Moses, 2001; see also Benson, Sabbagh, Carlson, & Zelazo, 2013; Wellman, 2014), development of the domain-specific knowledge required to succeed on a false belief task occurs from observation and interaction with the world (Wellman & Liu, 2004). Crucial to the discussion here, that domain-specific knowledge allows children to reason successfully on both the self and other questions equivalently.

Although the initial studies that asked children both the self and other questions found conflicting results,[1] strong evidence in favor of this view came from a meta-analysis of over 50 studies conducted by Wellman et al. (2001). This analysis showed that performance on the self and other questions did not significantly differ. Wellman et al. (2001) highlighted this finding as particularly contrastive to another theoretical account of theory of mind development – *simulation theory* – following research in philosophy of mind (e.g., Goldman, 2006; Gordon, 1992; see also Harris, 1992). Simulation theory suggests that inferences about others' mental states are made by accessing and reasoning about one's own mental states.

There are numerous philosophical descriptions of this approach to understanding theory of mind, including the possibility that the two theories reduce to one another (see e.g., Perner & Brandl, 2009), but the one specifically taken up by Wellman et al. regarded the primacy of one's own mental access (Harris, 1992). In particular, children's understanding of their own mental states is salient and necessary to understand similar mental states in others. Wellman et al. (2001), however, argue that the null result between performance on the self and other questions was meaningful for distinguishing between these two accounts: "… simulation accounts emphasizing the primacy of self-experience suggest that self understanding should develop first." And, later in on that page, they write, "on the surface, the fact that children ever systematically err in reporting their own false beliefs seems problematic for simulation accounts." (both p. 678).

Recent evidence, however, suggests that young children do have some first-person access to their mental states, particularly their knowledge states. Harris, Yang, and Cui (2017) demonstrated that 2- and 3-year-olds (children who typically do not succeed on the unexpected contents task) were more likely to deny possessing knowledge (i.e., saying "I don't know") than point out that another person does not know something (i.e., saying "You don't know"). These same children were also more likely to ask questions about whether others know something as opposed to whether they do. The former finding suggests that children might be aware of their own ignorance before that of others, while the latter suggests that children recognize their own lack of knowledge, and that others might be sources of knowledge that they themselves do not possess. These data suggest that although young children might not behave differently on self and other questions on a false belief task, they show differences in their understanding of their own and others' knowledge.

These data suggest the importance of reconsidering developmental findings on the difference between the self and other questions, and in particular the findings from the Wellman et al. (2001) analysis. In particular, while the difference between the self and the other questions was not significant, the interaction between the type of question asked and children's age was the level of a statistical trend ($p > .07$) with a small effect size ($f = 0.12$).[2] Typically, small effects are not investigated in developmental science because of limits on data collection or argumentation dismissing them as unimportant. However, small effects may be meaningful if they have important theoretical consequences (Ellis, 2010). Here, the theoretical implications of recent research suggesting children have first-person access to their mental states, warrants a reinvestigation of this finding. Doing so also allows us to address a methodological limitation of method used by Wellman et al. (2001), which is that their findings did not directly consider the sample size of each condition they analyzed.

The rest of this paper presents three analyses that reconsider and update the relation between performance on self and other questions about false belief through meta-analysis and a quasi-mega-analysis[3] of performance on the unexpected contents task. Analysis 1 presents a meta-analysis that follows a similar strategy used by Wellman et al. (2001). The goal of this analysis is not to replicate all aspects of Wellman et al. (2001), but rather focus specifically on the self-other difference. To that end, the analysis relies on papers that presented data from the unexpected contents task, which specifically asked either the self or other question. The papers collected for this analysis report how children performed on the self or other questions individually or how children performed on both questions (not combined into a single datum). Analysis 1 uses the same analysis strategy used by Wellman et al. (2001), but also uses a second measure of effect size, which takes the size of the samples into account to ensure the reproducibility of results.

A limitation of this first analysis is that some researchers who administer an unexpected contents task only ask the self question, while other researchers only ask the other question, while still other researchers ask both questions. Combining these between-subject data might not be as robust an analysis as one that contrasts results using a within-subject design. Analysis 2 replicates the metanalytic strategy on the subset of studies from Analysis 1 that present data from the self and other questions from the same participant.

Analysis 3 uses a similar within-subjects approach by considering a dataset of over 1200 preschoolers administered the unexpected contents task. These data were collected in a single lab with similar materials and the same procedure. This mega-analytic strategy not only generalizes and extends the results of the two meta-analyses, but also considers why small effects might be robust and meaningful, even if they are difficult to describe in individual empirical papers.

1. Analyses 1 and 2

The goal of these analyses is to examine whether there are differences in how children respond to questions about changes in their own previously erroneous belief states (self questions) and measures of others' possessing a false belief (other questions). We considered papers that reported the results of studies that use the unexpected contents task on children between the ages of 36 and 60 months, based on the assumption that the majority of children will eventually succeed on both questions at later ages. Analysis 1 uses a meta-analytic technique similar to that used by Wellman et al. (2001), updating the literature considered as well as extending the analysis to include a more robust measure of effect size. Analysis 2 then considers a subset of the conditions/experiments used in Analysis 1, which specifically contrast performance by the same child on both measures.

2. Literature review

To determine papers to analyze, the papers used in the Wellman et al. (2001), as well as a more recent analysis that exclusively considered studies reporting the unexpected contents task (Sobel & Austerweil,

[1] Perner et al. (1987) found that children performed better on the self than other question, whereas Gopnik and Gopnik and Astington (1988) found the opposite, that children performed better on the other question than the self question.

[2] The implicature of reporting $p > .07$ is that the significance level of this analysis is between $p = .07$ and $p = .08$. The effect size of this analysis was not reported, but was calculated based on the reported *F*-statistic.

[3] A mega-analysis is a reanalysis of raw data from multiple sources (Eisenhauer, 2021). Here, the raw data are all taken from a single lab, hence the "quasi."

2016) were examined. All the papers from those analyses that reported data from the unexpected contents task, which met the inclusion criteria (described below) were included. Google Scholar searches were also performed for the terms "unexpected content task," "representational change task" and "theory of mind scales". The first two are commonly used names for the measure. The latter was included because the scales (Wellman & Liu, 2004) present children with the unexpected contents task. The author also put out a call on the COGDEVSOC listserv (March 2021) for unpublished datasets and recently published data. Fig. 1 shows a flowchart of the literature review process. Because of the intention to include data from the author's published work separately in the mega-analysis presented in Study 3, none of the author's own papers were included in these studies (to reduce the possibility of bias in reproduction).

To summarize this process, after duplicate entries were removed from the initial search, there were 917 entries under consideration. One hundred fifty-two of those entries could be eliminated without review because (a) they were written by the author (8), (b) Google Scholar indicated they were books, book chapters, or published in journals that exclusively published review, and not empirical articles (88), and (c) Google Scholar indicated they were unavailable in English (56). Nine additional entries were removed because they were inaccessible.

The remaining 756 full-text articles and datasets were read by the author to determine whether the paper reported the age of an individual sample and performance on the self and other questions of the unexpected contents task individually (or if only one question was asked, performance on that question). Papers that only reported a composite score of performance on the self and other questions, or "false belief" scores that reported the results of an unexpected content task combined with other measures of false belief (such as the unexpected transfer task or the unexpected identity task) were not included unless the separate scores on the self and other questions of the unexpected contents task could be determined. Moreover, only samples of neurotypical children with a mean age between 36 and 60 months were considered. If the paper reported a longitudinal or training study, only pretest data was used; these kinds of studies were only included if they reported all pretest data as opposed to only including data from children who failed the measure. If the paper contrasted conditions in which deception was and was not used, only cases in which the motive of the study was not to deceive another person were considered (because there are well-known main effects of deceptive motive, see Wellman et al., 2001, and our supplemental materials). Fig. 1 shows the distribution of exclusions.

Following Wellman et al. (2001) and Glass, McGaw and Smith (1981), condition (or age group within a condition where possible) was used as the unit of analysis, instead of individual participant or a full study or paper. This resulted in including 91 papers/datasets used in Analysis 1, which represented 244 age groups/conditions within experiments (9201 data points in total). Analysis 2 focused on the subset of these 91 papers/datasets that asked both the self and other question of the same child participant to consider a within-subject analysis. This resulted in an analysis of the results of 37 papers/datasets (70 unique age groups/conditions, 2466 data points). The papers/datasets used in both analyses are reported in Table 1, as we as the number of conditions each entry contributed to the overall analysis.

3. Coding

Analyses 1 and 2 focused on extracting performance on the self and other question of the unexpected contents task, as well as the mean age and sample size of the reported sample. Each condition within each experiment was coded for performance on the test question as well as whether the measure was a self or other question and whether both questions were asked of the same child in the same procedure. If both questions were asked, we reported whether the order of these questions was counterbalanced. Finally, we only included whether the paper reported the mean age of the condition under consideration and the number of participants in that condition.

The goal of this analysis was not to reproduce all of the analyses performed by Wellman et al. (2001). Many of the analyses that were significant there, which might interact or affect the main result presented here, however, are described in the Supplemental Materials. This includes country of origin, whether a deceptive condition was included in the study, whether the deceptive object was really present in the container when the test question was asked, and whether temporal markers were used when asking the self question. One additional factor – the way in which researchers scored the task – was also considered in the Supplemental Materials, as previous work has shown that the developmental trajectory of children's performance is affected by how the measure is scored (Sobel & Austerweil, 2016).

To investigate differences between the self and other question on the unexpected contents measure, we conducted the same logit analysis as presented in Wellman et al. (2001). The proportion correct on each question (*p*) was transformed via a logit transformation in which the logit score = ln(*p*/1-*p*),[4] which became the dependent variable in the analyses. This analysis, however, does not take the sample size of the condition/age group into account. To account for this, we supplemented this analysis with a second transformation on the data, following the meta-analytic strategy used by Tong, Wang, and Danovitch (2020). Where possible (i.e., if the standard deviation of performance was reported or could be calculated), we transformed performance in each condition/age group to a Hedges' *g* value, based on the proportion correct and standard deviation of that condition/age group. This was done by assuming that the comparison value of chance had a Mean and Standard Deviation of 0.5 (an idealized representation of chance responding), via the formula:

$$g = \frac{(PropCorrect - 0.5)}{\sqrt{\frac{StdDev^2+.25}{2}}}$$

Two readers, blind to the hypotheses of the study coded a subset of 10 of the papers in the same manner. Agreement between these two readers was 100%. Agreement between these readers and the author was 90%. Disagreement was resolved through discussion.

4. Results

All data sets described in this manuscript (as well as a list of all articles that were considered) are available at https://osf.io/624av. We first consider the overall dataset (Analysis 1) and then a subset of the papers we considered, which presented both the self and other questions to the same child, using within-subject analyses (Analysis 2).

4.1. Analysis 1: Papers reporting findings from either self or other questions on the unexpected contents task

Fig. 2 shows the logit values for the self and other questions across age. A set of hierarchical linear regressions were conducted, with logit scores and *g*-values as the dependent variable. Each first considered age (defined by the mean age of the condition as reported in each paper) and question type (self vs. other). Then the interaction value between age and question type was added in a second model.

For the logit scores, the first model explained a significant amount of

[4] One condition presented a proportion of 0, while another presented a proportion of 1, both with standard deviations of 0. These resulted in an undefined logit scores and values of *g*. We entered proportions of 0.01 and 0.99 and standard deviations of 0.0001 for each, which set minimal and maximal values for the dependent variables without skewing the data too greatly. The results presented in the main text are similar if these two entries are not included in the analyses. Please note that the formula used is ln(*p*/1-*p*) and that the value is not raised to the fourth power.

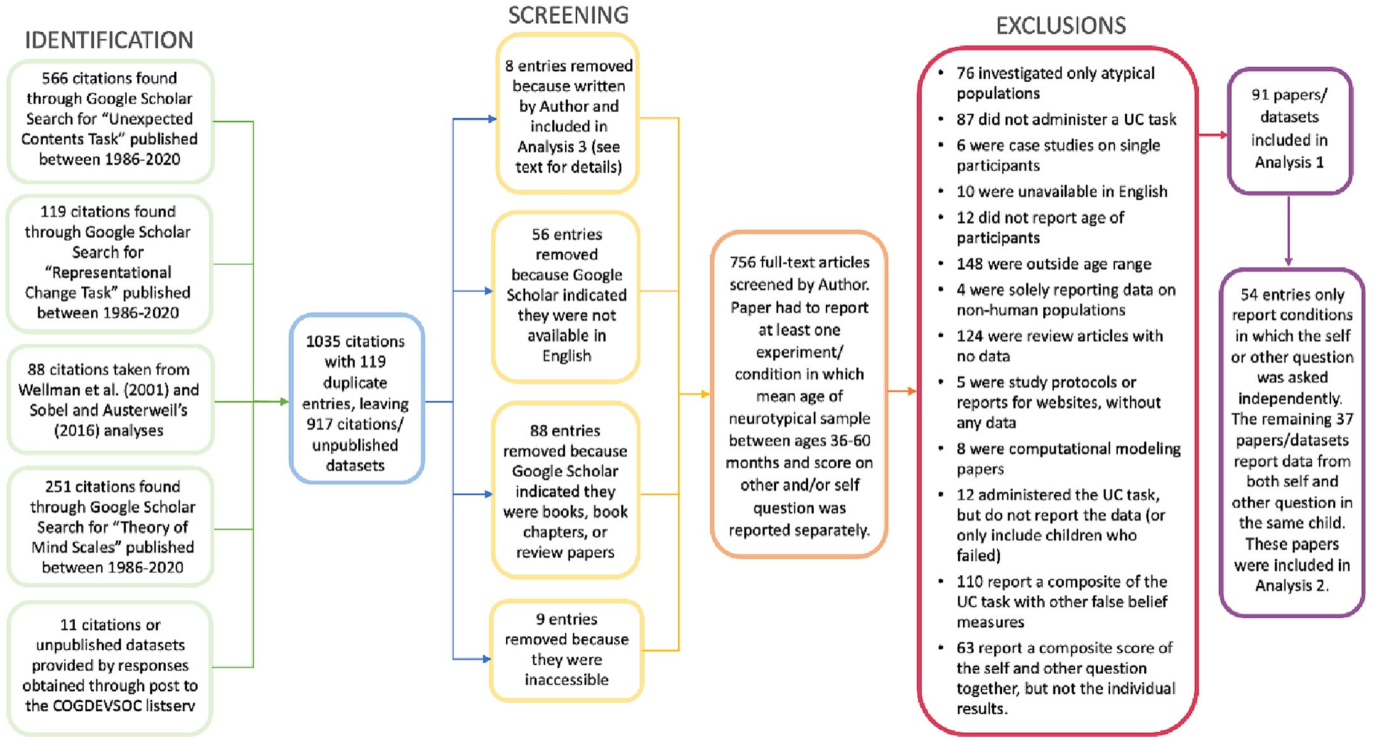

IDENTIFICATION
566 citations found through Google Scholar Search for "Unexpected Contents Task" published between 1986-2020
119 citations found through Google Scholar Search for "Representational Change Task" published between 1986-2020
88 citations taken from Wellman et al. (2001) and Sobel and Austerweil's (2016) analyses
251 citations found through Google Scholar Search for "Theory of Mind Scales" published between 1986-2020
11 citations or unpublished datasets provided by responses obtained through post to the COGDEVSOC listserv
1035 citations with 119 duplicate entries, leaving 917 citations/ unpublished datasets
SCREENING
8 entries removed because written by Author and included in Analysis 3 (see text for details)
56 entries removed because Google Scholar indicated they were not available in English
88 entries removed because Google Scholar indicated they were books, book chapters, or review papers
9 entries removed because they were inaccessible
756 full-text articles screened by Author. Paper had to report at least one experiment/ condition in which mean age of neurotypical sample between ages 36-60 months and score on other and/or self question was reported separately.
EXCLUSIONS
• 76 investigated only atypical populations
• 87 did not administer a UC task
• 6 were case studies on single participants
• 10 were unavailable in English
• 12 did not report age of participants
• 148 were outside age range
• 4 were solely reporting data on non-human populations
• 124 were review articles with no data
• 5 were study protocols or reports for websites, without any data
• 8 were computational modeling papers
• 12 administered the UC task, but do not report the data (or only include children who failed)
• 110 report a composite of the UC task with other false belief measures
• 63 report a composite score of the self and other question together, but not the individual results.
91 papers/ datasets included in Analysis 1
54 entries only report conditions in which the self or other question was asked independently. The remaining 37 papers/datasets report data from both self and other question in the same child. These papers were included in Analysis 2.

Fig. 1. Flowchart representing procedure used for literature identification, screening, and exclusions for Analyses 1–2.

the variance in the logit values over an intercept-only model, $R^2 = 0.28$, $F(2, 241) = 46.21$, $p < .001$. The second model, which added the interaction between age and question type, predicted a significant increase in the amount of variance in the logit values, $\Delta R^2 = 0.01$, $F(1, 240) = 4.07$, $p = .04$. This final model predicted a significant amount of variance over an intercept-only model, $R^2 = 0.29$, $F(3, 240) = 32.56$, $p < .001$. In this final model, age positively predicted performance, B = 0.07, SE = 0.02, 95% CI [0.04, 0.10], $t = 4.72$, $p < .001$. There was also a significant effect of question type, with performance on the self question higher overall than performance on the other question, B = 2.42, SE = 1.00, 95% CI [4.39, 0.45], $t = 2.42$, $p = .02$. There was also a significant interaction between question type and age, B = 0.04, SE = 0.02, 95% CI [0.00, 0.08], $t = 2.02$, $p = .04$.

A similar analysis for the *g*-values generated by each condition was conducted. The first model (age and question type), explained a significant amount of the variance over an intercept-only model, $R^2 = 0.30$, $F(2, 227) = 49.32$, $p < .001$. The second model, which added the interaction between age and question type, predicted a significant increase in the amount of variance in the logit values, $\Delta R^2 = 0.01$, $F(1, 226) = 4.23$, $p = .04$. This final model predicted a significant amount of variance over an intercept-only model, $R^2 = 0.32$, $F(3, 226) = 34.76$, $p < .001$. In this final model, age positively predicted performance, B = 0.03, SE = 0.01, 95% CI [0.02, 0.05], $t = 4.84$, $p < .001$, OR = 1.03. There was also a significant effect of question type, with performance on the other question lower overall than performance on the self question, B = 1.12, SE = 0.45, 95% CI [0.2.00, 0.23], $t = 2.50$, $p = .01$, OR = 0.33, and a significant interaction between question type and age, B = 0.02, SE = 0.01, 95% CI [0.00, 0.04], $t = 2.06$, $p = .04$. OR = 1.04.

4.2. Analysis 2: Papers reporting findings from both self and other questions on the unexpected contents measure

The second analysis looked at a subset of the papers from Analysis 1 that reported results of the self and other questions for the same children. Table 1 indicates the papers included in this analysis. As in the previous analysis, the logit value and the *g* value for performance on the self and other questions in each condition/age groups were considered. These data are shown in Fig. 3. A Generalized Estimating Equation was constructed to control for within-subject effects, with a robust correlation matrix, assuming a linear distribution on the dependent variables with age and question type as independent variables. For both analyses, a factorial model was considered.

Looking at the logit scores, the model revealed a significant effect of age, B = 0.12, SE = 0.02, 95% CI [0.08, 0.16], Wald $\chi^2(1) = 31.26$, $p < .001$, OR = 1.13, and question type, B = 2.91, SE = 0.87, 95% CI [1.21, 4.61], Wald $\chi^2(1) = 11.28$, $p = .001$, OR = 18.36, with children performing better overall on the self questions than the other questions. There was also a significant interaction between age and question type, B = 0.06, SE = 0.02, 95% CI [0.09, 0.02], Wald $\chi^2(1) = 11.25$, $p = .001$, OR = 0.94. As with the previous analysis, the younger children in this sample initially performed better on the self question than the other question, but as the age of children in the sample increased, performance on the other question started to exceed performance on the self question.

The same pattern of results was obtained when the Hedges' *g* scores were analyzed. There was a significant effect of age, B = 0.05, SE = 0.01, 95% CI [0.03, 0.07], Wald $\chi^2(1) = 34.70$, $p < .001$, OR = 1.05, and question type, B = 1.15, SE = 0.32, 95% CI [0.52, 1.78], Wald $\chi^2(1) = 12.91$, $p < .001$, OR = 3.16 as well as interaction between age and question type, B = 0.02, SE = 0.01, 95% CI [0.03, 0.01], Wald $\chi^2(1) = 10.82$, $p = .001$, OR = 0.98.

5. Discussion

Both the more inclusive analysis (Analysis 1) and the within-subject analysis (Analysis 2) of the self and other questions showed two findings of note. First, in addition to the expected main effects of age, there was a significant difference between performance on the self and other question, with children performing better on the former over the latter overall. The second was a significant interaction between children's age and question type, which potentially explains the first finding. Younger children performed better on the self question than the other question. However, as children increased in age from 36 to 60 months,

Table 1
Lfist of papers/Datasets used fin Anaflysfis 1 (and Anaflysfis 2, shown wfith an *).

Authors	Date of pubflficatfion	Number of age groups/ Condfitfions used fin Anaflysfis 1
Afltvater-Mackensen*	Unpubflfished	4
Amado, Serrat & Sfidera*	2014	2
Appfleton & Reddy*	1996	2
Arreckx	2007	1
Atance	2001	1
Atance, Metcaflf & Thfiessen	2017	2
Atance & ONefiflfl	2004	4
Bafird & Moses	2001	2
Barreto, Osorfio, Baptfista, Fearon, & Martfins*	2018	2
Befißert, Muflvey & Kfiflflen	Unpubflfished	2
Beflflagamba, Addessfi, Focaroflfi et afl.*	2015	4
Benson, Sabbagh, Carflson, & Zeflazo	2017	1
Bernstefin	2009	1
Bernstefin*	Unpubflfished	2
Bozbfiyfik	2016	1
CarflsonMoses*	2001	4
Cassfidy	1998	2
Causey & Bjorkflund*	2014	2
Conry-Murray	2013	1
Daflke	1995	2
Davfis*	2001	2
Fabrficfius & Khaflfifl	2003	1
Fflynn*	2006	4
Fflynn, O'Maflfley and Wood	2004	1
Freeman & Lacohee	1995	6
Frye, Zeflazo, and Paflfafi*	1995	8
Gopnfik & Astfington*	1988	8
Gopnfik & Rosatfi*	2001	2
Gopnfik & Sflaughter	1991	2
Guajardo, Parker & Turfley-Ames	2009	1
Guajardo & Turfley-Ames	2004	2
Gut, Haman, Gorbanfiuk, & Chyflfinskfia*	2020	4
Hafla, Hug & Henderson	2003	2
Hansen	2010	2
Hanson & Atance	2014	2
Hasnfi	2015	1
Hfiflfler, Weber & Young	2014	1
Hogrefe, Wfimmer, & Perner	1986	3
Hoflmes, Bflack & Mfiflfler*	1996	2
Hong*	2016	4
Hughes	1998	1
Jackson	2001	1
Juflfien	2018	1
Kaflfish, Wefisman & Bernstefin	2000	4
Kammermefier & Pauflus	2018	2
Keenan, Oflson and Marfinfi*	1998	4
Krachun, Carpenter, Caflfl, and Tomasefllo	2010	3
Kuntoro, Peterson & Sflaughter	2017	2
Lackner, Bowman & Sabbagh	2010	1
Lesflfie & Thafiss*	1992	3
Lewfis, Huang & Rooksby*	2006	4
Lewfis & Osborne*	1990	18
Lfiflflard	1993	3
Loke	2010	1
Mahy, Bernstefin, Gerrad & Atance*	2017	8
Major, Franco, and Zotovfic	2010	1
Metcaflf	2015	2
Mfitcheflfl & Lacohee	1991	2
Moflzhon	2016	1
Moore, Pure & Furrow*	1990	2
Moses & Fflaveflfl	1990	1
Muflfler, Mfiflfler, Mfichaflczyk, & Karapfinka	2007	2
Muflfler, Zeflazo, & Imrfisek*	2005	2
Nafito, Komatsu, and Fuke*	1994	4
Nawaz, Hanfif & Lewfis*	2015	4
Pecora, Addessfi, Paoflettfi & Beflflagamba*	2017	2
Perner, Leekam, & Wfimmer*	1987	2
Pesch, Semenov & Carflson	2020	1
Peterson & Sfiegafl	1999	1
Qufistberg	2018	1
Repachoflfi & Trapoflfinfi*	2004	2
Rfiggs and Robfinson	1995	2
Rfiggs, Robfinson & Samuefl	1996	4
Rubfio-Fernandex	2019	2
Ruffman, Oflson, Ash, and Keenan	1993	5
Sabbagh, Bowman, Evrafire, and Ito	2009	1
Safltmarsh, Mfitcheflfl, and Robfinson*	1995	4
Shehaefian, Nfieflsen, Peterson, & Sflaughter	2014	4
Sflaughter*	1998	2
Sflaughter & Gopnfik*	1996	4
Stanzfione	2015	1
Suflflfivan & Wfinner*	1993	4
Tardfif, Weflflman, & Cheung	2004	2
Tayflor & Carflson*	1997	4
Weflflman & Lfiu	2004	1
Wfiflflfiams & Happe*	2009	2
Wfiflflfiams & Happe*	2010	4
Wfimmer & Hartfl*	1991	7
Woflfle, McLaughflfin and Hefiphetz	2021	1
Yueng, Muflfler & Carpendafle*	2019	2
Zfiv, Zakafi-Mashfiach, Afl-Yagon, & Dromfi	2014	2

Notes. Aflfl pubflfished references are provfided fin the suppflementafl materfiafls sectfion.

performance on the other questfion caught up to, and then overtook performance on the seflf questfion.

Thfis reanaflysfis of the flfiterature on the unexpected contents task findficates that chfifldren's abfiflfity to finfer others' faflse beflfief and thefir own representatfionafl change have dfifferent deveflopmentafl trajectorfies. Before dfiscussfing the fimpflficatfions of these data, we want to consfider whether these ffindfings are aflso present fin a flarge-scafle dataset. Thfis wfiflfl provfide estfimates of the power of both the dfifference between the two types of questfions and the finteractfion between the questfions and age. Moreover, finvestfigatfing a flarge-scafle data set can provfide finsfight finto why these effects mfight have not prevfiousfly been descrfibed fin the flfiterature.

6. Analysis 3

To consfider the robustness of these ffindfings, these anaflyses were aflso consfidered on a dataset of preschoofl-aged chfifldren who partficfipated fin an unexpected contents procedure admfinfistered by the author's flab. These data were coflflected between 2001 and 2012. There was flfittfle varfiance fin experfimenter and materfiafls. Aflfl parents of chfifldren who partficfipated consented to the procedure, fin accordance wfith Brown Unfiversfity IRB (#0111991083, *Development of causalfity and fintentfionalfity*, #0503991803, *Chfildren's causal learnfing and developfing knowledge of mechanfisms* and #1007000219, Chfifldren's deveflopfing finferences from others' experfiences). Thfis dataset has been descrfibed fin a prevfious finvestfigatfion (Sobefl & Austerwefifl, 2016), but the anaflyses presented here are novefl.

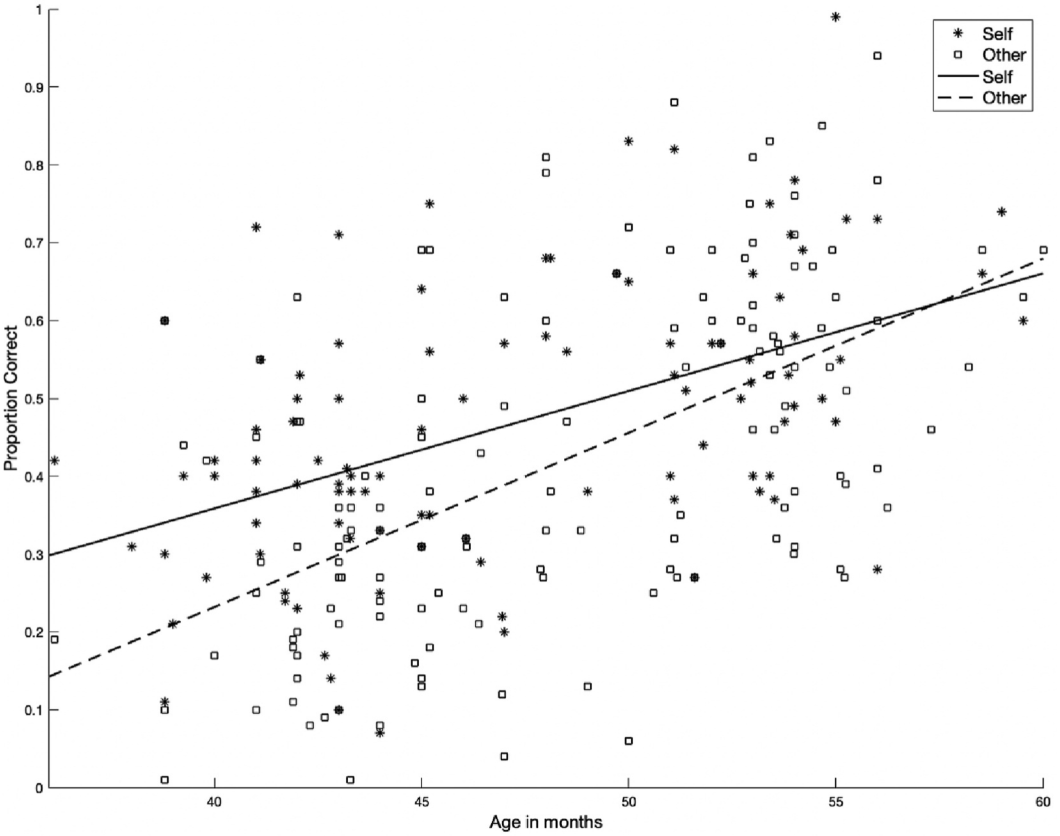


Fig. 2. Dfistrfibutfion of Performance on Representatfionafl Change (Seflf) and Faflse Beflfief (Other) Questfions fin Anaflysfis 1. Each pofint fin the graph represents one age group/condfitfion.

7. Method

7.1. Partficfipants

The ffinafl sampfle contafined the resuflts of 1231 chfifldren between the ages of 36–60 months (Mean Age = 50.90 months, SD = 6.08 months[5]). The sampfle was skewed towards oflder chfifldren wfith more 4-year-oflds fin the sampfle (*N* = 892, Mean age = 54.08 months, SD = 3.27) than 3-year-oflds fin the sampfle (*N* = 339, Mean age = 42.53 months, SD = 3.08 months). Three addfitfionafl chfifldren were tested fin thfis sampfle, but not fincfluded because they fafifled to respond to one of the three questfions used fin the procedure (see beflow).

7.1.1. Materfials

The majorfity of chfifldren were shown a deceptfive Crayofla crayons contafiner that contafined smaflfl bfirthday candfles shown fin Ffig. 4. A smaflfl number of chfifldren were gfiven the same procedure usfing a Band Afids box that contafined crayons.

7.1.2. Procedure

Chfifldren were seated across from the experfimenter at a tabfle. They were shown the cflosed deceptfive contafiner and asked what they thought was finsfide the box. Chfifldren typficaflfly responded approprfiatefly, (e.g., “crayons” or a sfimfiflar, pflausfibfle response, such as “markers” “coflors” “pencfifls”). Chfifldren were then shown the actuafl contents of the box (fin thfis case, the bfirthday candfles), whfich were taken out of the box and shown to the chfifld. These contents were then pflaced back finto the box and the box was cflosed.

The experfimenter then asked the *other* questfion fin whfich chfifldren were asked about the beflfief state of another person (a caregfiver flfike Daddy or a frfiend of the chfifld's, who had been mentfioned before the box was brought out, and who was not present durfing the tfime of the test). Specfifficaflfly, “Let's say <person> comes fin here. <Person> has never seen thfis box before. What wfiflfl<person> thfink fis fin the box? If the chfifld dfid not respond or safid “I don't know”, the experfimenter woufld ask the chfifld to make a guess.

After chfifldren generated thefir response to the other questfion, the experfimenter then asked the *self* questfion: “Before I showed you what was fin the box, what dfid you thfink was fin the box?” Agafin, fif the chfifld dfid not respond or safid, “I don't know”, the experfimenter woufld ask the chfifld to make a guess. Ffinaflfly, the experfimenter asked the *control* questfion: “What fis reaflfly fin the box?”

8. Results and discussion

Ffig. 5 shows performance on the seflf and other questfion fin the sampfle overaflfl, as wefllfl as the Spearman correflatfions wfith age. A Generaflfized Estfimatfing Equatfion was constructed to controfl for wfithfin-subject effects, assumfing a bfinomfiafl dfistrfibutfion for performance on each questfion. Age and questfion type were the findependent varfiabfles, assumfing a factorfiafl modefl. Thfis modefl reveafled a mafin effect of age, B = 0.12, SE = 0.01, 95% CI [0.10, 0.14], Wafld $\chi^2(1)$ = 120.65, $p <$ .001, OR = 1.12, and a mafin effect of questfion type, B = 1.63, SE = 0.67, 95% CI [0.31, 2.95], Wafld χ (1)^2 = 5.89, p = .02, OR = 5.11.

Note that unflfike Anaflyses 1–2, here, the overaflfl performance on the other questfion (55% accurate) outperformed performance on the seflf questfion (52% accurate). Thfis dfifference fis potentfiaflfly caused by the skew fin the age range of the sampfle as the finteractfion between age and questfion type was aflso sfignfifficant, B = 0.03, SE = 0.01, 95% CI [0.06,

[5] Unfortunatefly, the gender breakdown of the sampfle coufld not be reported because fin some cases, chfifldren's gender was not recorded.

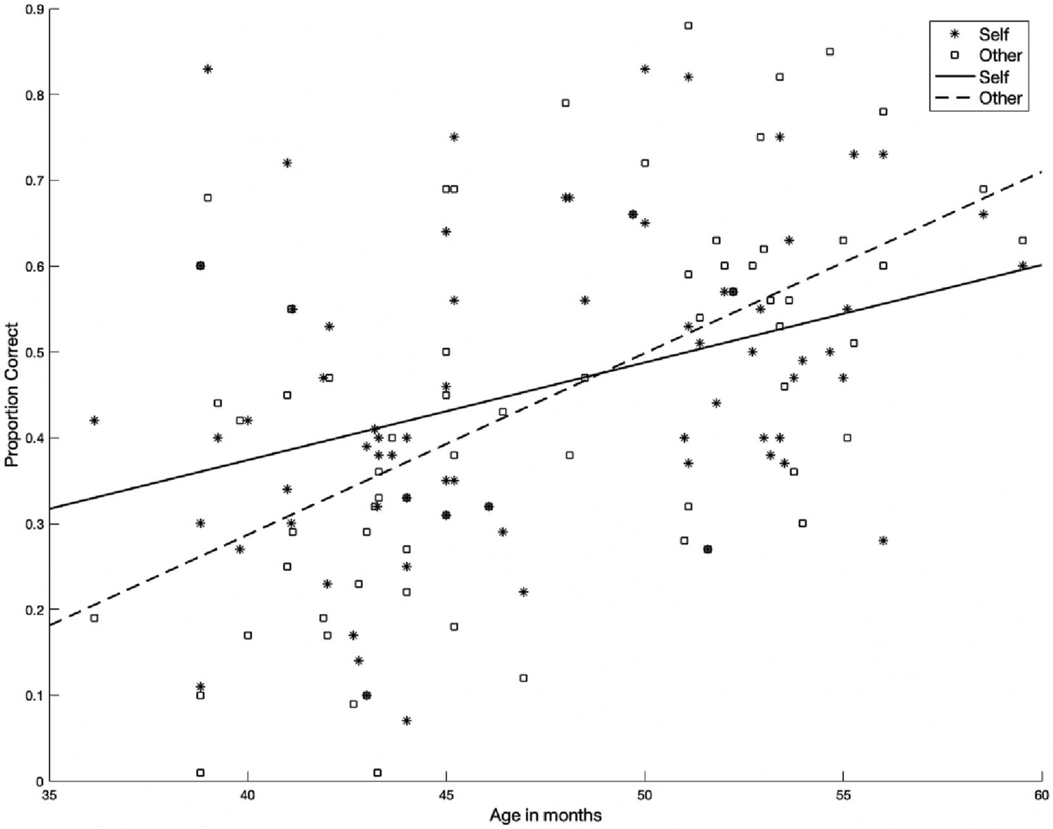


Fig. 3. Distribution of Performance on Representational Change (Self) and False Belief (Other) Questions in Analysis 2. Each point in the graph represents one age group/condition.

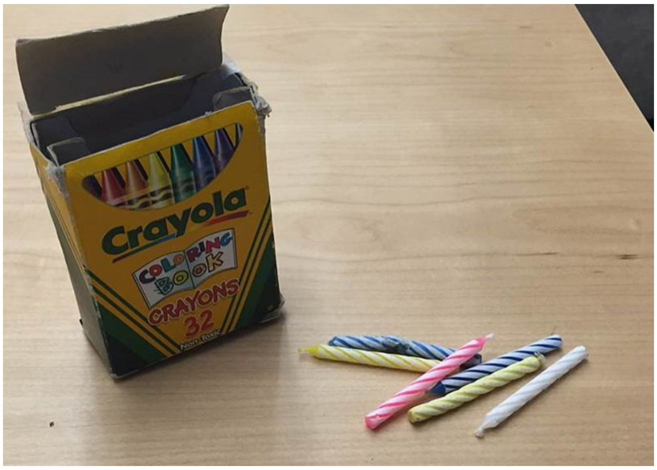


Fig. 4. Crayon box and candles used in unexpected contents task administered to children.

0.01], Wald $\chi^2(1) = 6.82$, $p = .009$, OR = 0.96. The mean age of the sample here is 50.90 months, compared with 47.89 months in Analysis 1 and 47.46 months in Analysis 2.

To consider the relation between the self and other questions, a binary logistic regression was conducted on performance on the other question with age and performance on the self question as predictors, as well as the interaction. The overall model was significant, $\chi^2(3) = 200.72$, $p < .001$, but only age was a significant predictor, B = 0.10, SE = 0.02, Wald $\chi^2(1) = 48.25$, $p < .001$, Odds Ratio = 1.11 95% CI [1.08, 1.14]. Neither performance on the self question, B = 1.20, SE = 1.12, Wald $\chi^2(1) = 1.15$, $p = .28$, or the interaction, B = 0.004, SE = 0.02, Wald $\chi^2(1) = 0.03$, $p = .86$, was significant. Removing the interaction from this analysis did not result in a significant increase in the model's predictive power, indicated by a change in 2 Log Likelihood = 0.03, $p = .86$. However, without the interaction term (which is how the relation between self and other has been analyzed), performance on the self question was now a significant predictor, B = 1.00, SE = 0.13, Wald $\chi^2(1) = 63.77$, $p < .001$, Odds Ratio = 2.72 95% CI [2.13, 3.48]. This further suggests the possibility that the self and other question have different developmental trajectories.

Analysis of this sample also sheds light on why this interaction may have been particularly difficult to uncover in previous studies using the unexpected contents task. A Monte-Carlo simulation was built, which constructed 10,000 samples of equal numbers of 3-year-olds and 4-year-olds of different sample sizes from this group of 1231 children. For each sample, the same GEE model described above was run to analyze the effects of age (in months), question type and their interaction. The mean p-values and the number of samples (out of 10,000) where $p < .05$ for the main effects of age and question type, and the age x question type interaction for each sample size are shown in Table 2. The effect of age was often detected in this simulation, while the effect of question type or the interaction would not often be detected. Even with large sample sizes ($N = 200$), those effects would be significant less than half the time. Given that the mean sample size used in Analysis 2 (the analysis in which we examined findings that compared performance on the self and other question from the same children) was N = ~35, it is possible that this finding has gone unnoticed in the literature.

A concern with the present analysis is that it might have revealed a simple explanation for the interaction we have observed. The unexpected contents task in Analysis 3 used a fixed question order, always asking the other question first and the self question second. It is possible that children simply switch their answer from one question to the other given the similarity between the other and self questions. This

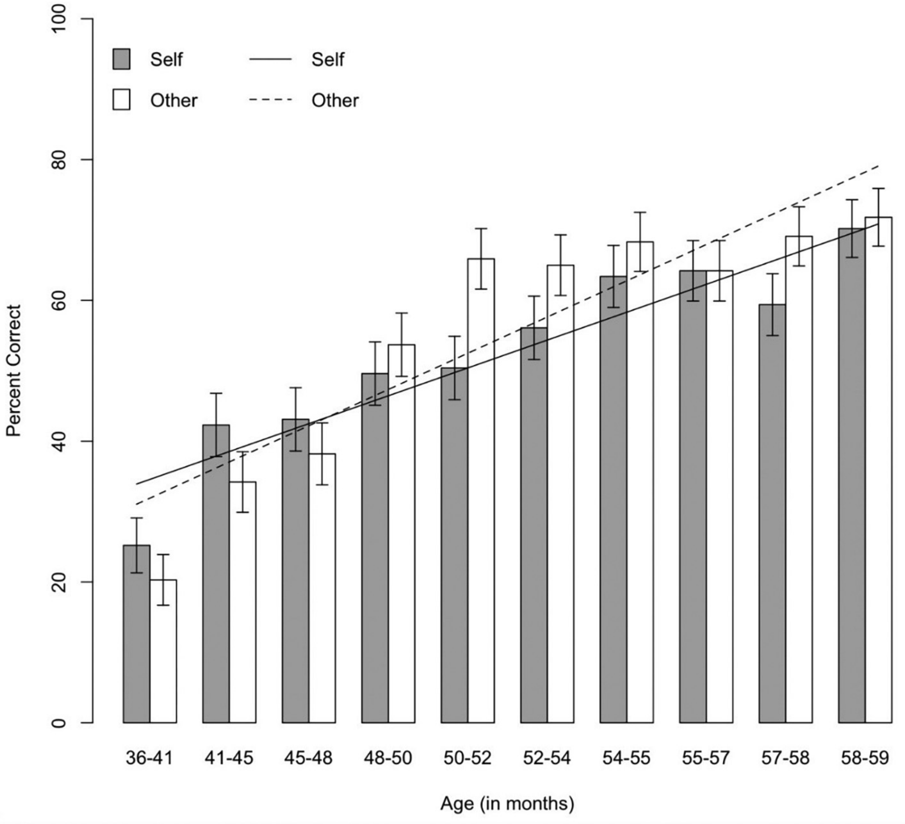


Fig. 5. Proportfion of chfifldren who responded to the seflf and other questfions correctfly across ages, dfivfidfing the sampfle by ten age percentfifles (Error bars show standard error). The soflfid and dashed flfines represent the best ffittfing Spearman correflatfions between performance and age for the seflf and other questfions respectfivefly.

Table 2
Probabfiflfity of detectfing sfignfifficant effects of age, questfion type, and finteractfion for sfimuflated experfiments wfith dfifferent sampfle sfizes (N = number of chfifldren fin each age group).

	Mean *p*-vaflue	Number of sampfles where *p* < .05 (out of 10,000)
Age, *N* = 30	0.04	8310
Age, *N* = 60	0.003	9907
Age, *N* = 100	0.00008	10,000
Age, N = 200	<0.00000001	10,000
Questfion Type, N = 30	0.43	1061
Questfion Type, N = 60	0.37	1501
Questfion Type, N = 100	0.29	2219
Questfion Type, N = 200	0.14	4319
Questfion Type x Age Interactfion, N = 30	0.43	1054
Questfion Type x Age Interactfion, N = 60	0.37	1513
Questfion Type x Age Interactfion, N = 100	0.29	2275
Questfion Type x Age Interactfion, N = 200	0.15	4373

expflanatfion seems unflfikefly, however, for severafl reasons. Ffirst, the flogfic of thfis expflanatfion fis that younger chfifldren were more flfikefly to respond fincorrectfly on the ffirst (other) questfion, and then swfitch thefir answer to the correct response on the second (seflf) questfion. As chfifldren get oflder, the flfikeflfihood that they respond correctfly on the ffirst questfion fincreases, hence the finteractfion. In the procedure used fin Anaflysfis 3 (sfimfiflar to aflfl papers consfidered fin Anaflysfis 2), chfifldren were gfiven no feedback to the response to the ffirst questfion. Whfifle chfifldren mfight finterpret thfis as findficatfing they shoufld swfitch thefir answer, fit fis equaflfly flfikefly that they finfer that because they dfid not get correctfive feedback, thefir response was correct and they shoufld respond fin the same manner. Moreover, chfifldren were not more flfikefly to swfitch thefir answer fif they responded correctfly to the ffirst questfion (36% of these chfifldren) than fif they dfid not (34% of these chfifldren), $\chi^2(1) = 0.45$, $p = .50$.

Second, even fif thfis expflanatfion hefld for the resuflts of Anaflysfis 3, the finteractfion was observed fin Anaflyses 1 and 2. In Anaflysfis 1, the data mostfly came from findependent sources (fi.e., condfitfions that onfly asked one of the seflf or other questfions). In Anaflysfis 2 (flfike Anaflysfis 3), both questfions were asked of the same chfifld, but dfifferent flabs used dfifferent methods for admfinfisterfing the procedure. Anaflysfis 2 was reexamfined to consfider how the seflf and other questfion were counterbaflanced fin these condfitfions. Thfirty-seven of the 70 condfitfions used a ffixed order *opposfite* to what was presented fin Anaflysfis 3 (fi.e., seflf questfion ffirst, other questfion second). Reanaflyzfing just those condfitfions demonstrated an fidentficafl pattern of resuflts to what was reported fin Anaflysfis 2 above, specfifficaflfly sfignfifficant mafin effects of age and questfion type, as wefll as a sfignfifficant finteractfion between age and questfion type.

9. General discussion

The three anaflyses presented here show a reproducfibfle finteractfion between preschooflers' performance on the seflf and other questfions of the unexpected contents measure and thefir age. Young 3-year-oflds were more accurate at keepfing track of thefir own representatfionafl change than finferrfing another's faflse beflfief, whfifle oflder 4-year-oflds show the reverse pattern. Thfis finteractfion produced a sfignfifficant dfifference between the two questfion types overaflfl, wfith performance on the seflf question exceeding that of performance on the other question in both Analysis 1 and 2. The reverse was true in Analysis 3, presumably due to the sample in Analysis 3 being older on average (the average ages in the first two analyses were both below 48 months, while the average age of the sample in the third analysis was over 50 months).

In each analysis, the size of the interaction effect was small, but consistent throughout the ways the data were considered. Given the small effect size, one might ask how these findings shed light on the way children develop inferences about false belief. In the next section, I try to discuss a possible interpretation and implications of this pattern of data.

10. Early privileged access, but later memory errors?

One possible interpretation of these data is that they support the possibility that children do have early privileged access to their own knowledge, as performance on the self question exceeds performance on the other question for the younger children examined in each analysis presented here. But one should not conclude from this (as potentially suggested by Piaget, 1932) that young children have a conceptual form of egocentrism, and simply used their privileged access to their own mental states for inferences about others. Rather, one might suggest the present analyses build on arguments made by Harris (2018), who suggested that children might have “privileged conscious access only to those beliefs and emotions that we assume to be based on a valid picture of reality. Once that picture of reality is shown to be mistaken, our beliefs and emotions will be revised and it will require an act of reconstruction to identify and explicate what we once thought and felt.” (p. 94). Critically, Harris's suggestion rested around children's understanding of their own ignorance, which develops earlier than their understanding of others' false belief (e.g., Pratt & Bryant, 1990; Wellman & Liu, 2004), but there is a similarity between knowing one is ignorant and knowing one is believing falsely. It seems likely that children might first appreciate their own ignorance, and then use that knowledge to learn that they are wrong about the world.

On this view, early in development, children do not understand that seeing candles in the crayon box requires them to change their belief about what is in the box. And given this, children might privilege their own beliefs about the actual contents, and retain that information over any inference they make about others' mental states (leading to better performance on the self question). Once, however, children appreciate that the container's contents being different from the expectations that are in common ground (i.e., the label on the box) requires a change in their own representation of the contents, the act of remembering their past belief might become more difficult than judging what another will believe. That is, the self question is no longer a question about children's own introspective access, but rather now becomes a memory question – specifically whether children remember their previous belief state (a point generally raised by Perner et al., 1987).

And when the self question is viewed as a memory task, there is a potential way to explain how performance on the other question exceeds it later in development. Children might make what Bernstein, Atance, Meltzoff, and Loftus (2007) call an error of “hindsight bias” (p. 1374). When presented with the deceptive contents, children err on the self question because they are overly influenced by what they now know to be true. Bernstein et al. showed that preschoolers' performance on different false belief tasks correlated with the amount of hindsight bias they displayed. These correlations were significant when the battery of false belief measures they administered had self questions; the battery in which only questions about others' false beliefs were asked did not show a significant relation to hindsight bias (see Table 3 on p. 1383). Moreover, there is a U-shaped developmental trajectory of hindsight bias; Bernstein, Erdfelder, Meltzoff, Peria, and Loftus (2011) found that although preschoolers (3- to 5-year-olds) engaged in more hindsight bias than older children, 3-year-olds were less biased than 4-year-olds (although these differences were either marginal trends or not statistically significant), and 5-year-olds also showed less bias than 4-year-olds. That is, 4-year-olds' greater hindsight bias suggests the possibility that the self question would be more difficult for them than the other question at this age, leading to their potentially performing better on the other question.

11. Developing altercentrism?

The possibility presented in the previous suggestion might be integrated with another, related literature in theory of mind development. Over the last 15 years, numerous studies have documented theory of mind capacities in children much younger than those under consideration here. A prototypical way that this literature has been represented is through examining toddlers' performance on measures of implicit false belief (e.g., Onishi & Baillargeon, 2005). These studies generally indicate that toddlers appreciate that others can have false beliefs.

The nature and reproducibility of the work on toddlers' implicit understanding of false belief is contentious, and beyond the scope of this discussion (see e.g., Sabbagh & Paulus, 2018 and the special issue of *Cognitive Development* they edited on this topic). Nonetheless, a promising theoretical approach was suggested by Southgate (2020). Early in development, infants register certain facets of the mental states of others – they have an altercentric bias (see also Apperly & Butterfill, 2009, for related arguments), but lack a clear sense of self. This explains their performance on certain early competence measures during the toddler years, as the presence of others can affect certain attentional cues (e.g., Manea, Kampis, Wiesmann, & Southgate, 2021). By the time children are 3, they begin to develop a fuller sense of self and represent events from their own perspective and that of others, but cannot link those representations together (following Perner & Leahy, 2016). And much like the arguments made above, their developing sense of self might make their own mental states more salient and available for reasoning (similar to the arguments made by Harris, 2018).

But as children begin to link those perspectives together and their capacities becomes representational, they do not simply a return to the altercentric bias they have in infancy. Rather what emerges is a focus on the importance of others' mental states, which could explain the better performance on the other question than the self question. Understanding others' mental states are more explanatory to understanding minds and engaging in social-cognitive inferences is consistent with suggestions made Gopnik (1993); see also Gopnik & Astington, 1988). The presence of others highlights the importance of those others' mental states, and attention to others' mental states begins to be more salient during the fifth year of life.

12. Implications, objections, and limitations

The present results could be interpreted as suggesting that the phenomenological experience of developing children in these samples is that they first have a growing awareness of their own mental states, potentially starting with ignorance and then moving towards their own false beliefs, and later in development, juxtaposed that privileged access to their mental states with the emergence of memory errors (that reduces performance on the self question) or the salience of others' mental states (that bolsters performance on the other question). However, the small effect sizes reported here suggest that the developmental story might not be as explicit.

So, it is also possible that what the present analyses suggest is a more minor change in the way researchers conceptualize the development of representational theory of mind. That is, close philosophical examination of how simulation theory would allow children to understand others' mental states might require them to replicate domain-specific knowledge indicated by positing children develop a representation theory of mind, indicating that there's not a great difference between those theories (e.g., Davies, 1994; Perner & Brandl, 2009).

And over the last few years, the idea that children develop such a representational theory of mind (i.e., theory theory) has evolved to posit that children possess both domain-general reasoning mechanisms and domain-specific knowledge (Gopnik & Wellman, 2012; Wellman, 2014). On this more contemporary view, children's domain-general causal learning capacities underlie their developing theory of mind, including their developing understanding of false belief. Children must possess the domain-specific knowledge necessary to engage in conceptual change, but are reliant on domain-general learning capacities for development to occur.

Framed in this way, mental state inferences from others' actions has been described as a form of "inverse planning;" reasoners reconstruct the mental states of others based on their actions, with numerous computational models describing adults' and infants' inferences (e.g., Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Rafferty, LaMar, & Griffiths, 2015). This framework suggests that reasoners generate and represent possible models of how mental states cause actions. Forming and reasoning over such a hypothesis space might be a domain-general mechanism related to children's imaginative abilities (e.g., Kushnir, 2022; Sobel, in press), and prior to appreciating the specific representational nature of belief, might make one's own mental states – i.e., the basis of such generative models – more salient. This would allow children the kind of privileged access under consideration early in development. But another reasoning capacity that is developing between the ages of 3–4 is the capacity to posit that nonobvious causes are efficacious (Nazzi & Gopnik, 2000; Sobel, Yoachim, Gopnik, Meltzoff, & Blumenthal, 2007). This suggests that when children realize that beliefs – that is, nonobvious causes of behaviors – are representational, their attention switches from the self to the other. This could allow better performance on the other question than the self question later on.

Put another way, the small effect sizes documented here might not indicate that children have broad-sweeping different phenomenological experiences when engaging in false belief reasoning during the preschool years. Rather, these findings might indicate smaller demand characteristics of a broader, domain-general reasoning system. Future research should investigate this possibility to shed light on these theoretical contrasts.

Finally, a limitation of the present discussion is the need to integrate the developmental findings presented here with findings in adult social psychology that emphasize the similarities and distinctions between mental state inference about the self and other. In particular, these studies suggest that mental state inferences about others are made by accessing a shared neural representation of our own and others' actions, thoughts, and emotions, and selecting the information that specifically applies to others and not ourselves (e.g., Decety & Sommerville, 2003; Epley, Keysar, Van Boven, & Gilovich, 2004; Gallese & Goldman, 1998; Jenkins, Macrae, & Mitchell, 2008; Preston and De Waal, 2002; Steinbeis & Singer, 2014; although see Saxe, 2005, 2009, for a dissenting interpretation of these data). To illustrate how such an integration might be considered, the interaction between age and question type described here presumably does not extend to adulthood. Adults and even older children would presumably perform at ceiling on both question types (see supplemental materials section for an extension of the mega-analysis including older children). It is certainly possible that the interaction reported here has no bearing on adults' use of simulation-like capacities to take the perspective of others. However, it is also possible that these results indicate an early-developing attentional focus on other people, as evident by numerous cases in which adults' social (and even non-social) inferences are influenced by the presence of others (see e.g., Kampis & Southgate, 2020, for a review).

13. Methodological implications

It is also necessary to describe an important methodological implication of the present analyses. Many studies considered for the analysis could not be included because they reported the results of a composite score of the self and other questions (63 papers) or combined such a composite score with other measures of false belief (110 papers). An open question is whether the results of those studies are actually measuring a composite false belief measure or show different patterns of results if the self and other questions are analyzed separately. An important methodological implication of the present analysis is for researchers to justify this practice, while recognizing that the self and other questions might not be measuring the same construct (see Supplemental Materials for one such analysis).

14. Conclusion

The present analyses reveal a mostly hitherto neglected aspect[6] of the development of false belief on the unexpected contents task – an interaction between children's age between the third and fifth birthdays and their ability to answer questions about their own representational change and others' false belief correctly. These results are not easily explainable simply by positing that children have privileged access to their own mental states (which would not posit the shift to better performance on the other question later in development), nor by classic accounts of children developing a representational theory of mind (which posits no difference between these questions). I have tried to speculate on a novel interpretation based on existing literature, as well as refinements to contemporary accounts of theory of mind development. But these speculations require further investigation and refinement. That said, the broad conclusion is to challenge researchers moving forward to revise theories of the development of false belief to include the possibility that performance on the self and other questions have a different, but potentially related developmental trajectory.

Credit author statement

David Sobel (Brown University) is the sole author of this manuscript. He conceptualized, analyzed, and wrote the contents of this manuscript.

Data availability

All data sets described in this manuscript (as well as a list of all articles that were considered) are available at https://osf.io/624av

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105403.

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*, 953. https://doi.org/10.1037/a0016923

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour, 1*(4), 0064. https://doi.org/10.1038/s41562-017-0064 |

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Benson, J. E., Sabbagh, M. A., Carlson, S. M., & Zelazo, P. D. (2013). Individual differences in executive functioning predict preschoolers' improvement from theory-of-mind training. *Developmental Psychology, 49*(9), 1615–1627. https://doi.org/10.1037/a0031056.

[6] To be fair, the interaction between the self and other questions and age has been reported previously in one paper (Mitchell & Neal, 2005). This paper, however, did not meet our criteria for inclusion (it focuses on an unexpected transfer task). Moreover, age in that paper is analyzed categorically, and the categories appear to be determined arbitrarily (42- to 51-month olds vs. 54- to 65-month-olds, with the former group containing many fewer children than the latter). Thus, while the interaction described here is not unprecedented, the findings reported here are more robust.

Bernstefin, D. M., Atance, C., Mefltzoff, A. N., & Loftus, G. R. (2007). Hfindsfight bfias and deveflopfing theorfies of mfind. *Chfild Development, 78*, 1374–1394. https://dofi.org/10.1111/j.1467-8624.2007.01071.x

Bernstefin, D. M., Erdfeflder, E., Mefltzoff, A. N., Perfia, W., & Loftus, G. R. (2011). Hfindsfight bfias from 3 to 95 years of age. *Journal of Experfimental Psychology: Learnfing, Memory, and Cognfitfion, 37*, 378. https://dofi.org/10.1037/a0021971

Carflson, S. M., & Moses, L. J. (2001). Indfivfiduafl dfifferences fin finhfibfitory controfl and chfifldren's theory of mfind. *Chfild Development, 72*, 1032–1053. https://dofi.org/10.1111/1467-8624.00333

Davfies, M. (1994). The mentafl sfimuflatfion debate. *Phfilosophfical. Issues, 5*, 189–218. https://dofi.org/10.2307/1522880

Decety, J., & Sommervfiflfle, J. A. (2003). Shared representatfions between seflf and other: A socfiafl cognfitfive neuroscfience vfiew. *Trends fin Cognfitfive Scfiences, 7*(12), 527–533. https://dofi.org/10.1016/j.tfics.2003.10.004

Efisenhauer, J. G. (2021). Meta-anaflysfis and mega-anaflysfis: A sfimpfle fintroductfion. *Teachfing Statfistfics, 43*(1), 21–27. https://dofi.org/10.1111/test.12242

Eflflfis, P. (2010). *The essentfial gufide to effect sfizes: Statfistfical power, meta-analysfis, and the finterpretatfion of research results*. New York: Cambrfidge Unfiversfity Press.

Epfley, N., Keysar, B., Van Boven, L., & Gfiflovfich, T. (2004). Perspectfive takfing as egocentrfic anchorfing and adjustment. *Journal of Personalfity and Socfial Psychology, 87*(3), 327–339. https://dofi.org/10.1037/0022-3514.87.3.327

Fflaveflfl, J. H. (1999). Cognfitfive deveflopment: Chfifldren's knowfledge about the mfind. *Annual Revfiew of Psychology, 50*, 21–45. https://dofi.org/10.1146/annurev.psych.50.1.21

Gaflflese, V., & Gofldman, A. (1998). Mfirror neurons and the sfimuflatfion theory of mfind-readfing. *Trends fin Cognfitfive Scfiences, 2*(12), 493–501. https://dofi.org/10.1016/S1364-6613(98)01262-5

Gflass, G. V., McGaw, B., & Smfith, M. L. (1981). *Meta-analysfis fin socfial research*. Beverfly Hfiflfls, CA: Sage Pubflficatfions Inc.

Gofldman, A. I. (2006). *Sfimulatfing mfinds: The phfilosophy, psychology, and neuroscfience of mfindreadfing*. New York: Oxford Unfiversfity Press.

Gopnfik, A. (1993). How we know our mfinds: The fiflflusfion of ffirst-person knowfledge of fintentfionaflfity. *Behavfioral and Brafin Scfiences, 16*, 1–14. https://dofi.org/10.1017/S0140525X00028636

Gopnfik, A., & Astfington, J. W. (1988). Chfifldren's understandfing of representatfionafl change and fits reflatfion to the understandfing of faflse beflfief and the appearance-reaflfity dfistfinctfion. *Chfild Development, 59*, 26–37. https://dofi.org/10.2307/1130386

Gopnfik, A., & Mefltzoff, A. N. (1997). *Words, thoughts, and theorfies*. MIT Press.

Gopnfik, A., & Weflflman, H. M. (1994). The theory theory. In L. A. Hfirschfefld, & S. A. Geflman (Eds.), *Mappfing the mfind: Domafin specfiffficfity fin cognfitfion and culture* (pp. 257–293). Cambrfidge Unfiversfity Press.

Gopnfik, A., & Weflflman, H. M. (2012). Reconstructfing constructfivfism: causafl modefls, Bayesfian fleamfing mechanfisms, and the theory theory. *Psychologfical bulletfin, 138*(6), 1085–1108. https://dofi.org/10.1037/a0028044

Gordon, R. M. (1992). The sfimuflatfion theory: Objectfions and mfisconceptfions. In *, 7. Mfind and language* (pp. 11–34). https://dofi.org/10.1111/j.1468-0017.1992.tb00195.x

Harrfis, P. L. (1992). From sfimuflatfion to foflk psychoflogy: The case for deveflopment. *Mfind & Language, 7*(1–2), 120–144. https://dofi.org/10.1111/j.1468-0017.1992.tb00201.x

Harrfis, P. L. (2018). Revfisfing prfivfifleged access. In J. Proust, & M. Fortfier (Eds.), *Metacognfitfive dfiversfity: An finterdfiscfiplfinary approach* (pp. 83–96). New York: Oxford Unfiversfity Press. https://dofi.org/10.1093/oso/9780198789710.001.0001.

Harrfis, P. L., Yang, B., & Cufi, Y. (2017). 'I Don't know': Chfifldren's earfly taflk about knowfledge. *Mfind & Language, 32*, 283–307. https://dofi.org/10.1111/mfifla.12143

Jara-Ettfinger, J., Gweon, H., Schuflz, L. E., & Tenenbaum, J. B. (2016). The naïve utfiflfity caflcuflus: Computatfionafl prfincfipfles underflyfing commonsense psychoflogy. *Trends fin Cognfitfive Scfiences, 20*(8), 589–604. https://dofi.org/10.1016/j.tfics.2016.05.011

Jenkfins, A. C., Macrae, C. N., & Mfitcheflfl, J. P. (2008). Repetfitfion suppressfion of ventromedfiafl prefrontafl actfivfity durfing judgments of seflf and others. *Proceedfings of the Natfional Academy of Scfience U S A, 105*, 4507–4512. https://dofi.org/10.1073/pnas.0708785105

Kampfis, D., & Southgate, V. (2020). Afltercentrfic cognfitfion: How others finffluence our cognfitfive processfing. *Trends fin Cognfitfive Scfiences, 24*(11), 945–959. https://dofi.org/10.1016/j.tfics.2020.09.003

Kushnfir, T. (2022). Imagfinatfion and socfiafl cognfitfion fin chfifldhood. *Wfiley finterdfiscfiplfinary revfiews*. *Cognfitfive Scfience*. https://dofi.org/10.1002/wcs.1603. e1603.

Manea, V., Kampfis, D., Wfiesmann, C. G., & Southgate, V. (2021). Testfing the afltercentrfism hypothesfis fin young finfants. In *Paper presented at the 2021 bfiennfial meetfing of the socfiety for research fin chfild development, vfirtual*.

Mfitcheflfl, R. W., & Neafl, M. (2005). Chfifldren's understandfing of thefir own and others' mentafl states. Part B. understandfing of others precedes seflf-understandfing for some faflse beflfiefs. *Brfitfish Journal of Developmental Psychology, 23*, 201–209. https://dofi.org/10.1348/026151004X21099

Nazzfi, T., & Gopnfik, A. (2000). A shfift fin chfifldren's use of perceptuafl and causafl cues to categorfizatfion. *Developmental Scfience, 3*(4), 389–396. https://dofi.org/10.1111/1467-7687.00133

Onfishfi, K. H., & Bafiflflargeon, R. (2005). Do 15-month-ofld finfants understand faflse beflfiefs? *Scfience, 308*(5719), 255–258. https://dofi.org/10.1126/scfience.1107621

Perner, J. (1988). Deveflopfing semantfics for theorfies of mfind: From proposfitfionafl attfitudes. In J. W. Astfington, P. L. Harrfis, & D. R. Oflson (Eds.), *Developfing theorfies of mfind*. Cambrfidge Unfiversfity Press. pp. 141.

Perner, J. (1991). *Understandfing the representatfional mfind*. Cambrfidge, MA: The MIT Press.

Perner, J. (1996). Sfimuflatfion as expflficfitatfion of predficatfion-fimpflficfit knowfledge about the mfind: Arguments for a sfimuflatfion-theory mfix. In P. Carruthers, & P. K. Smfith (Eds.), *Theorfies of theorfies of mfind*. Cambrfidge Unfiversfity Press. https://dofi.org/10.1017/CBO9780511597985. pp. 90.

Perner, J., & Brandfl, J. L. (2009). Sfimuflatfion à fla Gofldman: pretend and coflflapse. *Phfilosophfical Studfies, 144*, 435–446. https://dofi.org/10.1007/s11098-009-9356-z

Perner, J., & Leahy, B. (2016). Mentafl ffifles fin deveflopment: Duafl namfing, faflse beflfief, fidentfity and fintensfionaflfity. *Revfiew of Phfilosophy and Psychology, 7*, 491–508. https://dofi.org/10.1007/s13164-015-0235-6

Perner, J., Leekam, S. R., & Wfimmer, H. (1987). Three-year-oflds' dfiffficuflty wfith faflse beflfief: The case for a conceptuafl deffiicfit. *Brfitfish Journal of Developmental Psychology, 5*, 125–137. https://dofi.org/10.1111/j.2044-835X.1987.tb01048.x

Pfiaget, J. (1932). *The moral judgment of the chfild*. London: Routfledge.

Pratt, C., & Bryant, P. (1990). Young chfifldren understand that flookfing fleads to knowfing (so flong as they are flookfing finto a sfingfle barrefl). *Chfild Development, 61*(4), 973–982. https://dofi.org/10.1111/j.1467-8624.1990.tb02835.x

Premack, D., & Woodruff, G. (1978). Does the chfimpanzee have a theory of mfind? *Behavfioral and Brafin Scfiences, 1*, 515–526. https://dofi.org/10.1017/S0140525X00076512

Preston, S. D., & De Waafl, F. B. (2002). Empathy: Its ufltfimate and proxfimate bases. *Behavfioral and Brafin Scfiences, 25*(1), 1–20. https://dofi.org/10.1017/S0140525X02000018

Rafferty, A. N., LaMar, M. M., & Grfifffiths, T. L. (2015). Inferrfing flearners' knowfledge from thefir actfions. *Cognfitfive Scfience, 39*(3), 584–618. https://dofi.org/10.1111/cogs.12157

Sabbagh, M. A., & Pauflus, M. (2018). Repflficatfion studfies of fimpflficfit faflse beflfief wfith finfants and toddflers. *Cognfitfive Development, 46*, 1–3. https://dofi.org/10.1016/j.cogdev.2018.07.003

Saxe, R. (2005). Agafinst sfimuflatfion: The argument from error. *Trends fin Cognfitfive Scfiences, 9*(4), 174–179. https://dofi.org/10.1016/j.tfics.2005.01.012

Saxe, R. (2009). The neurafl evfidence for sfimuflatfion fis weaker than I thfink you thfink fit fis. *Phfilosophfical Studfies, 144*(3), 447–456. https://dofi.org/10.1007/s11098-009-9353-2

Sobefl, D.M. (fin press). Understandfing pretense as causafl finference. Deveflopmentafl Revfiew.

Sobefl, D. M., & Austerwefifl, J. L. (2016). Codfing chofices affect the anaflyses of a faflse beflfief measure. *Cognfitfive Development, 40*, 9–23. https://dofi.org/10.1016/j.cogdev.2016.08.002

Sobefl, D. M., Yoachfim, C. M., Gopnfik, A., Mefltzoff, A. N., & Bflumenthafl, E. J. (2007). The bflficket wfithfin: Preschooflers' finferences about finsfides and causes. *Journal of Cognfitfion and Development, 8*(2), 159–182. https://dofi.org/10.1080/15248370701202356

Southgate, V. (2020). Are finfants afltercentrfic? The other and the seflf fin earfly socfiafl cognfitfion. *Psychologfical Revfiew, 127*, 505–523. https://dofi.org/10.1037/rev0000182

Stefinbefis, N., & Sfinger, T. (2014). Projectfing my envy onto you: Neurocognfitfive mechanfisms of an offflfine emotfionafl egocentrficfity bfias. *NeuroImage, 102*, 370–380. https://dofi.org/10.1016/j.neurofimage.2014.08.007

Tong, Y., Wang, F., & Danovfitch, J. (2020). The rofle of epfistemfic and socfiafl characterfistfics fin chfifldren's seflectfive trust: Three meta-anaflyses. *Developmental Scfience, 23*, Artficfle e12895. https://dofi.org/10.1111/desc.12895

Weflflman, H. M. (1990). *The chfild's theory of mfind*. Cambrfidge, MA: The MIT Press.

Weflflman, H. M. (2014). *Makfing mfinds: How theory of mfind develops*. New York: Oxford Unfiversfity Press.

Weflflman, H. M., Cross, D., & Watson, J. (2001). Meta-anaflysfis of theory-of-mfind deveflopment: The truth about faflse beflfief. *Chfild Development, 72*, 655–684. https://dofi.org/10.1111/1467-8624.00304

Weflflman, H. M., & Lfiu, D. (2004). Scaflfing of theory-of-mfind tasks. *Chfild Development, 75*(2), 523–541. https://dofi.org/10.1111/j.1467-8624.2004.00691.x