Cognitive Development 40 (2016) 9–23

Contents lists available at ScienceDirect

Cognitive Development

Coding choices affect the analyses of a false belief measure[☆]

David M. Sobel[*], Joseph L. Austerweil

Brown University, United States

ARTICLE INFO

Article history:
Received 17 June 2015
Received in revised form 15 July 2016
Accepted 8 August 2016

Keywords:
Unexpected contents task
Theory of mind
Metaanalysis

ABSTRACT

The unexpected contents task is a ubiquitous measure of false belief. Not only has this measure been used to study children's developing knowledge of belief, it has impacted the study of atypical development, education, and many other facets of cognitive development. Based on a review of articles using this task, we show that there is no consensus regarding how to score this measure. Further, examining both a logit analysis of performance on this measure and performance of a large sample of preschoolers, we show that which coding scheme researchers used to analyze raw data from this measure has a reliable effect on results, particularly when smaller sample sizes are used. Integrating our results, we conclude that the most frequently used coding scheme is flawed. We recommend best practices for scoring the unexpected contents task, and that researchers examine how they analyze data from this measure to ensure the robustness of their effects.

Contents

[☆] This research was supported by a grant from the National Science Foundation (NSF 1223777). We would like to thank Rose DeRenzio, Isobel Heck, and Antonija Kolobaric who helped with coding, and Nadia Chernyak, Christopher Erb, Tamar Kushnir, Deanna Macris, and Tiffany Tassin for providing fruitful discussion on this topic. Joseph Austerweil is now at the University of Wisconsin-Madison.

[*] Corresponding author at: Department of Cognitive, Linguistic, and Psychological Sciences, Box 1821, Brown University, Providence, RI 02912, United States.

E-mail address: Dave_Sobel@brown.edu (D.M. Sobel).

http://dx.doi.org/10.1016/j.cogdev.2016.08.002
0885-2014/© 2016 Elsevier Inc. All rights reserved.

1. Introduction

For over 25 years, researchers have been interested in children's developing *theory of mind* – their developing knowledge of others' mental states (e.g., Flavell, 1999). While these investigations emphasize the nuanced and staggered developmental trajectory of children's mental state knowledge, one aspect of theory of mind development – children's understanding of another's false beliefs – has dominated both the theoretical and empirical landscape (see e.g., Apperly & Butterfill, 2009 Perner, 1991; Wellman, Cross, & Watson, 2001).

Measures of false belief typically fall into one of two categories, *unexpected transfer* and *unexpected contents.* In the unexpected transfer task (Wimmer & Perner, 1983), children are introduced to an agent, who is shown to have a belief about the location of an object, because s/he has perceptual access to that location. The object's location is then changed, and critically, the agent does not have perceptual access to or communication about the new location. Children are then asked about the (false) belief the agent has about the location of the object.

The unexpected contents task is similar in that children reason about another's false belief, but also are often probed about their own belief state. Children are introduced to a familiar container, which is revealed to be deceptive. For instance, in Perner, Leekham, and Wimmer (1987), children are introduced to a *Smarties* box (a type of candy), which is revealed to contain either a pencil or a pencil and Smarties while their friend is outside of the room. In their procedure, children are asked three questions after the contents of the box are placed back inside: A question regarding their memory: "Can you remember what's inside here?" (*control* question), a question asking about their knowledge of their own mental states (*self* question): "But what did you think was in here?", and a question asking about their knowledge of another's beliefs about the contents of the box (*other* question), "What will <name of friend> think is in here?" The results of studies using these procedures are highly consistent. In a metaanalysis of over 50 studies, Wellman et al. (2001) showed that children's ability to answer the test questions develops between the ages of 36 and 60 months in a manner that reflects their own understanding of representational change and that another can hold a false belief.

We have observed, however, that papers reporting this measure score the unexpected contents task differently. For example, Perner et al. (1987) reports the results of the test questions, but excludes the data generated by children who failed the control measure. They write that three of their participants "failed to remember there was a pencil. . .in the box (control question). Their responses to the test questions were therefore meaningless." (p. 133). Dissimilarly, Astington and Jenkins ((1999) see also Wellman and Liu (2004)) used a coding scheme in which "passing" the measure requires a correct response to the control question – that is, if children fail the control measure, they are counted as failing the task, regardless of their response to the test question(s). Many other papers either do not ask this memory question or do not consider the answer to this question when scoring the measure.

Here, we examine the role of coding scheme in research on children's developing false belief. We first examine how the unexpected contents task is scored. We review a subset of the literature that uses this measure, describing the different ways it has been scored. Next, we consider whether differences in the way this task is scored affects results and potential interpretations of findings. We conduct a logit analysis using methods similar to Wellman et al. (2001) to demonstrate that the choice of coding scheme overall influences analyses of performance. We then discuss a published article, which presented their results in sufficient detail such that we can calculate whether the statistical significance of their results depend on the choice of coding scheme. We find that the results in this paper might depend on the choice of coding scheme. We then analyze an in-house data set of ~1200 preschoolers who have been given this measure, using a bootstrap analysis to consider the potential effect of coding scheme on effect sizes. All of our analyses support the same conclusions: (1) There is no agreement regarding how this measure is scored; (2) the choice of coding scheme affects analyses, and (3) many researchers use what we believe is the least optimal coding scheme. Finally, to conclude the paper, we offer a set of best practices based on the results of our analyses.

2. Literature review

The goal of the literature review is to examine whether there is agreement in how researchers have scored the unexpected contents measure over the last 25+ years. We conducted a literature review on papers that report children's performance on an unexpected contents task and coded the manner in which data collected from children were scored.

To determine papers to analyze, we first examined the well-known metaanalysis on children's false belief knowledge (Wellman et al., 2001), and included every paper that reported data from the unexpected contents task. We then performed a Google Scholar search for the terms "unexpected content task" to find papers published between 2001 and 2013. Overall, we found 88 papers that reported at least one experiment using the unexpected contents task or a variant of the measure (123 separate experiments on 9365 participants). Table 1 reports these papers.

For each experiment, the first author and a research assistant (who was blind to the goals of this review) read the methods and scoring sections of the experiment to determine whether a control question regarding the child's memory of the contents of the box was asked. If so, the coders examined whether and how the control question was included in the scoring of the task. Experiments were categorized into eight distinct scoring methods:

Table 1
List of Papers used in the Literature Review.

Authors	Date of Publication
Appleton and Reddy	1996
Astington and Jenkins	1999
Atance and O'Neill	2004
Atance, Bernstein, and Meltzoff	2010
Baron-Cohen	1991
Bartsch and Wellman	1989
Benson, Sabbagh, Carlson, and Zelazo	2013
Bernstein, Atance, Meltzoff, and Loftus	2007
Bialystok and Senman	2004
Blair and Razza	2007
Blair, Granger, and Razza	2005
Carlson and Moses	2001
Carlson, Mandell, and Williams	2004
Carlson, Moses, and Claxton	2004
Cassidy, Fineberg, Brown, and Perkins	2005
Cassidy, Werner, Rourke, and Zubernis	2003
Cheung, Hsuan, Creed, Ng, Wang, and Lo	2004
Courtin	2000
Dalke	1995
De Villiers and Pyers	2002
Farrar and Maag	2002
Fisher, Happe, and Dunn	2005
Flynn	2006
Flynn, O'Malley, and Wood	2004
Freeman and Lacohée	1995
Frye, Zelazo, and Palfai	1995
Geren, Sneckder, and Shafto	Unpublished (Written 2009)
Gopnik and Astington	1988
Guajardo and Turley-Ames	2004
Guajardo and Watson	2002
Hala, Hug, and Henderson	2003
Hale and Tager-Flusberg	2003
Hale and Tager-Flusberg	2005
Hogrefe, Wimmer, and Perner	1986
Hughes	1998
Hughes and Cutting	1999
Hughes, Jaffee, Happe, Taylor, Caspi, and Moffitt	2005
Jackson	2001
Jenkins and Astington	1996
Kalish, Weissman, and Bernstein	2000
Kelley, Paul, Fein, and Naigles	2006
Krachun, Carpenter, Call, and Tomasello	2009
Lackner, Bowman, and Sabbagh	2010
Lackner, Sabbagh, Hallinan, Liu, and Holden	2012
Lalonde and Chandler	1995
Leslie and Thaiss	1992
Lewis and Osborne	1990
Lind and Bowler	2009
Lohmann and Tomasello	2003
Low, Goddard, and Melser	2009
Lundy	2002
Major, Franco, and Zotovic	2010
Mathews, Dissanayake, and Pratt	2003
Mitchell and Lacohée	1991
Moore, Pure, and Furrow	1990
Moses and Flavell	1990
Müller, Miller, Michalczyk, and Karapinka	2007
Müller, Zelazo, and Imrisek	2005
Naito, Komatsu, and Fuke	1994
Pellicano	2010
Pellicano, Mayberry, and Durkin	2005
Perner, Frith, Leslie, and Leekam	1989
Perner, Leekam, and Wimmer	1987
Peterson	2000
Peterson and Siegal	1999
Peterson, Wellman, & Liu	2005
Repacholi and Trapolini	2004
Robinson, Riggs, and Samuel	1996
Ruffman, Olson, Ash, and Keenan	1993
Ruffman, Slade, Rowlandson, Rumsey, and Garnham	2003
Sabbagh, Bowman, Evraire, and Ito	2009
Saltmarsh, Mitchell, and Robinson	1995
Slade and Ruffman	2005

Table 1 (*Continued*)

Authors	Date of Publication
Slaughter and Gopnik	1996
Slaughter, Dennis, and Pritchard	2002
Smith, Apperly, and White	2003
Symons, Peterson, Slaughter, Roche, and Doyle	2005
Tardif, Wellman, and Cheung	2004
Taylor and Carlson	1997
Taylor, Lussier, and Maring	2003
Tine and Lucariello	2012
Vinden	1996
Walker	2005
Watson, Nixon, Wilson, and Capage	1999
Watson, Painter, and Bornstein	2001
Williams and Happé	2009
Wimmer and Hartl	1991
Zelazo, Jacques, Burack, and Frye	2002

Notes. All references provided in the reference list.

Table 2
Distribution of Coding Schemes across the 92 Experiments.

Coding scheme Description	Number of Experiments in Sample	Overall Percentage
(1) Children asked a control question about the actual contents of the box. Children who answer incorrectly are not included in analyses.	12	13.04
(2) Children are asked a control question about the actual content of the box. Children who answer incorrectly are counted as not passing the measure.	24	26.09
(3) Children are asked a control question about the actual content of the box. Results of the control question are not factored into the scoring. Percentage of children who fail the control question is reported.	3	3.26
(4) Children are asked a control question about the actual content of the box. Results of the control question are not factored into the scoring. Percentage of children who fail the control question is not reported.	7	7.60
(5) No control question is asked.	38	41.30
(6) Children are asked a control question about the actual content of the box. Authors report that all children pass the control question.	6	6.52
(7) Results of control question are equated with results of test questions	1	1.09
(8) No coding scheme is provided	1	1.09

1) Participants who failed the control question were excluded from the analysis.
2) Participants who failed the control question were counted as failing the measure (or received a score of zero).
3) The control question was not factored into the score children received, but the number of children who failed the control was reported.
4) The control question was not factored into the score children received, and the number of children who failed the control was not reported.
5) No control question was asked.
6) Authors reported that all children passed the control.
7) The control question counts as a “correct” answer, equivalently to the test questions.
8) No coding scheme is reported.

Agreement on how the measure was scored was 90% with disagreements resolved through discussion between the two coders with one exception: For one experiment, the two coders could not come to agreement regarding a resolution. To resolve this, we contacted the corresponding author of the paper to ask questions that would allow us to code it appropriately.

If a paper reported multiple experiments, we examined whether different coding schemes were used across the experiments. If all the experiments had the same coding scheme, the paper was counted once (to ensure that a paper with four experiments was not weighed four times as much as a paper with only one experiment). If experiments had different coding schemes within the same paper, we counted each of those experiments. Four papers met this latter criterion, thus there were 92 experiments in the analysis.

2.1. Results

Table 2 shows the distribution of coding schemes used to analyze this measure across these experiments. Of these 92 experiments, six reported that the researchers administered a control question, which all children passed (Scoring Method 6 above). These papers do not report what researchers would have done had a child failed. One experiment counted the control question equivalently to the test questions (e.g., Scoring Method 7; children got a point for answering the “self”, “other”, and

Table 3
Distributions of Coding Schemes Based on Purpose of Study.

Studies Examining Children's Understanding of Belief		
Coding scheme Description	Number of Experiments in Sample	Overall Percentage
(1) Children who fail control are not included in analyses (Exclude System)	5	14.29
(2) Children who fail control are counted as not passing measure (Failure System)	8	22.86
(3) Control Question is not factored into scoring (Ignore System)	22	62.86
Total	35	
Studies Correlating False Belief to Other Measures		
(1) Children who fail control are not included in analyses (Exclude System)	7	14.28
(2) Children who fail control are counted as not passing measure (Failure System)	16	32.65
(3) Control Question is not factored into scoring (Ignore System)	26	53.06
Total	49	53.06

"control" questions correctly, for a score of 0–3 on the measure). One experiment did not report a coding scheme.[1] (Scoring Method 8). We did not consider these experiments further. This left us with 84 experiments to analyze.

The remaining coding schemes described above can be further categorized into three types. Coding Scheme 1 (Henceforth, *Exclude*): Exclude children who fail the control from the analyses (Scoring Method 1, represented by 12/84 experiments). Coding Scheme 2 (Henceforth, *Failure*): Treat children who fail the control as failing the measure (Scoring Method 2, represented by 24/84 experiments). Coding Scheme 3 (Henceforth, *Ignore*): Ignore the control question in the scoring (Scoring Methods 3–5, represented by 48/84 experiments). The distribution of these three coding schemes differed from a uniform distribution, $\chi^2(2, N = 84) = 24.00$, $p < 0.01$, $\phi = 0.53$. The Ignore scheme was the most popular, and was used significantly more often than either of the other coding schemes, Binomial tests, both p-values < 0.01.

The unexpected contents task may be conducted for a variety of reasons, and these reasons might motivate which coding scheme is used. Two of the main reasons include (1) A measure of theory of mind ability to correlate with performance on other measures, and (2) to study some facet of children's understanding of belief. We might expect that researchers engaged in the former type of study would be less likely to eliminate data based on a failure to answer the control question. In these cases, the false belief task is one measure among a large set that might involve a substantial investment of time and resources to collect, so attempts are made to retain as much data as possible. In contrast, the latter cases are often single procedures, where it might be easier to replace participants to achieve a large-enough sample.

To consider this question, we analyzed whether the goal of the experiment was primarily to relate performance on the false belief measure with another task or to understand the nature of belief. We coded 35 of the 84 experiments we analyzed here as being about the nature of belief and the remaining 49 experiments as being correlational in nature. Table 3 shows the distribution of these two subsets, and there was not a significant difference between the distributions, $\chi^2(2, N = 84) = 1.03$, $p = 0.60$, $\phi = 0.11$. This analysis suggests that concerns about data retention did not necessarily motivate the use of a particular coding scheme.

We next examined whether there were any differences among papers that used the three coding schemes. We first analyzed the sample size of the experiments, using nonparametric analyses because of unequal variance among the groups. A median test showed that the median sample size differed among experiments using the three coding schemes differs, $\chi^2(2, N = 84) = 9.00$, $p = 0.01$, $\phi = 0.33$. Post-hoc analysis revealed that the only significant difference among the sample sizes was the Failure scheme and the Ignore scheme, Mann-Whitney $U = 335.00$, $z = -2.88$, $p < 0.01$, $r = 0.34$. When researchers have fewer participants, they are more likely to use the coding scheme that does not factor performance on the control question into account. We also ran this same analysis on the Google Scholar citation count[2] and the age of the publication for the three kinds of coding schemes (cases where we would not expect any differences). There was no effect of coding scheme on either of these factors, $\chi^2(2, N = 84) = 0.00$ and 4.65, $p = 1.00$ and 0.10, $\phi = 0.00$ and $\phi = 0.24$, respectively.

2.2. Discussion

There was little agreement regarding how researchers scored the unexpected contents measure. The most common coding scheme was to not include the control question (if it was asked at all) into the coding scheme. The most infrequently used scheme was to exclude children who failed the control question from the data analysis.

Although we have found that researchers are more likely use the *Ignore* coding scheme than other coding schemes, we have not yet analyzed whether the differences among the coding schemes matter. We address this question in three ways. First, we conduct a logit analysis on the literature reviewed in the previous section. Second, we examine one particular published paper for differences in scoring systems. Third, since 2001, the first author and members of his lab have administered a

[1] An attempt was made to contact the corresponding author of the paper, but this individual did not respond

[2] Note that this analysis was run on May 4, 2015, so it is possible that these numbers have changed as citation counts are dynamic.

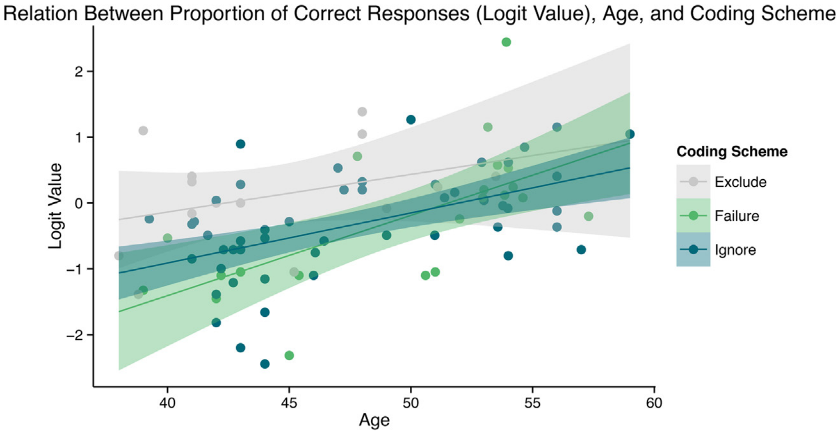

Fig. 1. Relations among coding scheme, age, and the proportion of correct responses on the unexpected contents task (transformed via the logistic function) given by children over the results of 86 conditions included in the analysis. Note that the Exclude Coding Scheme reported larger proportion of children passing the false belief task across all ages.

version of the unexpected contents measure to most of the children who have participated in studies. In many cases, this measure was used as a warm-up procedure to familiarize the child with the experimenter, and was irrelevant to the particular study. We compiled these data together for analysis below.

3. Logit analysis

To investigate whether there is an effect of the scoring system researchers used for the unexpected contents measure, we conducted the same logit analysis as presented in Wellman et al. (2001) using a subset of the papers included in the analysis above. We used condition (or age group within a condition where possible) as the unit of analysis, instead of individual participant or full study. This particular approach is useful because it analyzes performance on the measure in question, and does not require analysis of the frequency statistics presented in most published papers (e.g., see Glass, McGraw & Smith, 1981). That said, there are some limitations to this method, which we discuss below. However, as our goal was to investigate the impact scoring system had on established findings, we thought the best way to begin was to follow and extend a well-known analysis on false belief performance.

To construct our sample, we analyzed every experiment in each paper where we could determine performance correct on the unexpected content measure. We focused only on conditions that included typically developing children tested using a standard version of the unexpected contents test. If the paper reported a longitudinal or training study, we only used pretest data. Unlike the analysis reported in the previous section, we considered all the experiments in each paper and each condition or age group in these experiments, as long as we could recover the participants' mean age[3] and performance on the test questions. We only included conditions where the participants' mean age was between 36 and 59 months (inclusive). This resulted in our analysis including 86 separate conditions from 55 experiments among 40 papers. The total number of participants in this analysis was 2924. Five of the conditions reported here used a battery of theory of mind measures, in which several unexpected contents tasks were combined with one appearance-reality test, but one that used the same false belief and control questions. We included these in the analysis reported. Excluding them does not change the significance levels reported in the main text.

Each condition was coded for the proportion of correct responses to the test questions (the self and/or other questions). We also combined data from multiple unexpected contents trials, if administered to get a single proportion correct score. We recorded the scoring system used to score the unexpected contents task (as derived from the analysis in the previous section). Following Wellman et al. (2001), proportion correct (p) was transformed via a logit transformation in which the logit score = $ln(p/1\text{-}p)$[4] Fig. 1 shows the relation between age and logit score for each of the three scoring systems.

We first considered whether there was a general relation between age and logit score, which there was, Adjusted $R^2 = 0.20$, $F(1, 84) = 22.29$, $p < 0.001$. The effect size of this relation was computed in terms of an odds ratio, which was 2.36 per year (or 1.07 per month). The age at which children are ~50% correct is 51 months of age. An effect size of 2.36 per year means that there is an increase in probability of passing the unexpected contents task from ~50% at 51 to ~70% at 63 months (or

[3] In two cases, some children tested in the study were not given the unexpected contents measure, so the mean age of the sample does not specifically represent the mean age of the children given this measure. We used the reported mean age in the analysis. Excluding these two cases does not change the significance levels reported in the main text.

[4] One addition condition, not included in the analysis, presented a proportion of 1, which results in an incomputable logit score.

Table 4
Reconstruction of Data Set from Peterson et al. (2005), Analyzed with Different Coding Schemes.

Distribution using Failure Coding Scheme (as used in Peterson et al., 2005)			
	Native Deaf Signers (N = 11)	Late Deaf Signers (N = 36)	Children with Autism (N = 36)
Passed Unexpected Contents	9	12	17
Failed Unexpected Contents	2	24	19
$x^2_{Yates}(2, N = 83) = 6.17$, $p = 0.045$, $\varphi = 0.27$			
Distribution using Exclude Coding Scheme (Based on their report of performance on control question)			
Passed Unexpected Contents	9	12	17
Failed Unexpected Contents	2	19	16
$x^2_{Yates}(2, N = 75) = 4.45$, $p = 0.11$, $\varphi = 0.24$			

Note. Yates Corrections were used in this analysis because at least one cell has fewer than 5 entries, which potentially overestimates significance. See text for details.

from 50% at 51 months to 52% at 52 months). The effect size reported here is smaller, but similar to Wellman et al. (2001), who reported an odds ratio of 2.94 per year for the effect of age.

We then considered a hierarchical regression examining first age and then coding scheme to isolate the variance explained by coding scheme beyond the effects of age. The overall model predicted a significant amount of the variance in performance, $F(3, 82) = 11.88$, $p < 0.005$ and critically, scoring system predicted a significant amount of variance of performance, beyond the effect of age, Adjusted $\Delta R^2 = 0.077$, $F(2, 82) = 5.48$, $p = 0.006$. In this analysis, the Exclude scheme generally produced higher logit values from the Failure and Ignore schemes together, $b = -0.25$, SE = 0.077 (CI = [−0.40, −0.10]), $p = 0.001$, while Failure and Ignore did not differ from one another, $b = 0.039$ (CI = [−0.15, 0.23]), SE = 0.098, $p = 0.69$.

A difference between our analysis and the Wellman et al's analysis is that the average age of children at the ~50% performance mark was ~51 and ~44 months, respectively. There are several potential reasons for this difference. First, Wellman et al.'s analysis included three different types of false belief measures, while we focused on only one, and it is possible that the unexpected content measure was slightly more difficult. These different measures might also interact with this kind of analysis differently. Second, Wellman et al.'s analysis involved many older papers. The median age of papers using the Exclude scheme was 15 years old, which was the highest of the three coding systems. Fig. 1 shows that for the Exclude scheme, the mean age for 50% performance (logit score = 0) is ~42 months, which is quite similar to Wellman et al.'s (2001) analysis. It is possible that more papers published before 2001 used the Exclude scheme, which could account for some of the difference. It is also interesting to note that the Exclude scheme resulted in results most similar to Wellman et al's analysis.

The general conclusion of this logit analysis is that the coding scheme used by a researcher can have an influential role in the false belief scores from the unexpected contents measure. A limitation of this work (and the metaanalysis reported in Wellman et al., 2001) is that the sample sizes of the conditions are not taken into account as a way of normalizing these results (with sample size of conditions ranging from 10 to 128). We did rerun the analyses weighing the value of each logit by sample size, and achieved similar results in terms of effect sizes (e.g., the effect of scoring system beyond age in this analysis is Adjusted $\Delta R^2 = 0.102$). As such, we believe that the logit analysis that is presented provides evidence that choice of scoring system affects what level of performance is reported in a potential sample. We turn now to a question that follows from this analysis, which is whether the choice of coding scheme influences published results.

4. Analysis of Peterson, Wellman and Liu (2005)

In our literature review, we found one study where it was possible to reconstruct enough of the dataset, so that we could contrast at least two of the Coding Schemes presented above (Peterson et al., 2005), verified by Personal Communication with the first author, May 29, 2015). The analysis of this study is informative towards suggesting that choice of coding scheme for analysis may affect significance levels presented in results.

This study compared theory of mind development among children born deaf to deaf parents, those born deaf to hearing parents, and those diagnosed with autism. The unexpected contents measure was administered as part of the "Theory of Mind Scales" (Wellman & Liu, 2004), and was coded in the same manner as reported in that paper (using the Failure scheme).

One aspect of their analyses (reported on p. 507) was a test of performance on the unexpected contents measure among these three groups. This was done to compare their results with other papers that had investigated theory of mind performance in other samples of deaf children and children with autism. They reported a significant difference among the target groups, which motivated several subsequent analyses. Table 4 shows this data set and the analyses we present. We replicated one of their analyses (using a Yates correction, because one of the cells has fewer than 5 entries, making a standard

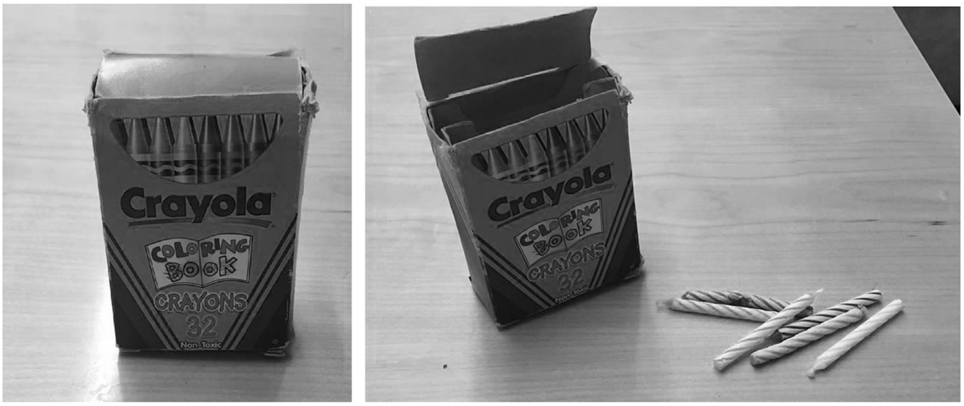

Fig. 2. The crayons-candles box used with most of the participants reported in the In-house data analysis.

Chi-Squared analysis less valid).[5] However, when the Exclude scheme is used instead of the Failure scheme, the analysis is no longer significant at an alpha level of 0.05.

Critically, we offer this analysis not as a criticism of their article or by means to question their results or the conclusions these researchers make. In fact, we commend them for providing the reader with enough information to reconstruct their full dataset, and of all the analyses they presented, this was the only one we found that was affected by choice of Coding Scheme. Using the Exclude instead of the Failure scheme results in the elimination of eight data points, which reduces the power of the analysis. Given that their article analyzed data from atypical populations, it may be appropriate to include all potential data points. Moreover, given that the Failure scheme potentially differentiates children who pass a measure from children who do not (the goal of administering this measure in their paper), we do not doubt the validity of their findings. Rather, we present this analysis as further evidence that the choice of coding scheme can influence whether a result is considered significant. We will return to this discussion in the section on best practices.

5. In-house data analysis

In this section, we examine a large dataset from children who participated in the same unexpected contents procedure. These data were collected between 2001 and 2012 by the first author and a set of students. In some cases, these data were reported in the published paper that emanated from the child's visit to the laboratory (Sobel, 2004, 2006, 2015; Sobel, Sommerville, Travers, Blumenthal, & Stoddard, 2009; Van Reet, Green & Sobel, 2015). In other cases, the data from the child's visit was reported, but the child's participation in the false belief measure was not reported, because it was irrelevant to the paper or because it was removed during the review process.. In other cases, the project did not result in a publication. The procedure was part of all relevant Institutional Review Board protocols and all parents of children who participated consented to the procedure, in accordance with the Institutional Review Board of the authors' university.

5.1. Method

5.1.1. Participants

The main sample analyzed here contains the data of 1231 children between the ages of 36 and 59 months (M_{age} = 50.90 months, SD = 6.08 months). There were fewer 3-year-olds (N = 339) than 4-year-olds (N = 892) in the data set. Data from 118 other children were also available. Three children failed to provide responses to one of the test questions. The other 115 were excluded because these children were outside of this age range (these data were mostly from 5- to 7-year-olds). Because we were concerned mostly with development during the preschool ages and because there were fewer children in the older age ranges, we chose to exclude these data so that we did not skew the main analysis. This decision also paralleled the logit analysis presented above.

5.1.2. Materials

The majority of children were shown a deceptive Crayola crayons container that contained candles shown in Fig. 2. A small number of children were given the same procedure using a Band Aids box that contained crayons.

[5] Our choice to correct these analyses is controversial. Some believe that a Yates correction should never be used because it overcorrects. Others suggest that they should only be used for 2 × 2 tables. Yates (1934) discusses the usage of his correction for 2 × 3 analyses, and suggests that it might be less necessary than for 2 × 2 analyses, but not unnecessary. That said, more recent statistical texts (e.g., Howell, 2006) suggest using a Fisher's Exact test in this circumstance. In this case, Coding Scheme 2 results in a p-value < 0.05 whereas Coding Scheme 1 is significant at exactly p = 0.05.

5.1.3. Procedure

Children were seated across from the experimenter, shown the deceptive container and asked what they thought was inside the box. Children typically responded “crayons” or a similar, plausible response (e.g., “markers” “colors” “pencils”).[6] Children were then shown the actual contents of the box (the birthday candles), and the candles were removed from the box so that children could see them up close. The candles were then placed back into the box and the box was closed. The experimenter then asked a *false belief other* question in which they were asked about the belief state of another person (a caretaker like Daddy or a friend of the child’s, who had been mentioned before the box was brought out, and who was not present during the time of the test). Specifically, “Let’s say <person> comes in here. <Person> has never seen this box before. What will <person> think is in the box? If the child did not respond or said ‘I don’t know’, the experimenter would ask the child to make a guess.

After children generated their response to the false belief other question, the experimenter then asked the *false belief self* question: “Before I showed you what was in the box, what did you think was in the box?” Again, if the child did not respond or said, “I don’t know”, the experimenter would ask the child to make a guess. Finally, the experimenter asked the *control* question: “What is really in the box?” The 1231 children included in this analysis all provided responses to all three questions.

5.2. Results

5.2.1. Analysis of overall data set

We first report the overall level of performance on each of the three questions. Children responded correctly on the false belief other question 55% of the time (678 out of 1231 children), correctly on the false belief self question 52% (645/1231) of the time, and correctly on the control question 88% (1087/1231) of the time. This last finding is consistent with the majority of published reports on false belief. For instance, Wellman et al. (2001) only included studies in their metaanalysis that yielded 80% performance on control questions or better. The present data are thus consistent with other previously published studies.

To assess whether using different coding schemes could yield significantly different results, we calculated the false-belief scores according to the different coding schemes for each participant. Children received a score of 1 if they answered the false belief other question correctly and a score of 1 if they answered the false belief self question correctly. We analyzed these data via three coding schemes under discussion. The Exclude scheme was the sum of the score on false belief self + false belief other, excluding those participants who failed false belief control. The Failure scheme was the sum of false belief self + false belief other for participants who passed the false belief control and zero for those who failed the false belief control. The Ignore scheme was the sum of false belief self + false belief other regardless of whether the participant passed or failed the false belief control question. Fig. 3 depicts performance by age for each of the three coding schemes.

We examined the effect of coding scheme using a series of multi-level models. We used both multi-level models where the age of the child is treated as a continuous variable (months-old; denoted *age*) and where it is treated as a categorical variable (3-year-old vs. 4-year-old; denoted *age group*) because both methods for analyzing experiments are used in the literature. We also performed post-hoc contrasts and analyzed the main effects and interactions for the age × coding scheme and age group × coding scheme models. When compared to a baseline model, a model only containing coding scheme had a significant main effect, $\chi^2(2) = 141.14$, $p < 0.001$. Adding either age or age group had a significant main effect $\chi^2(1) = 172.55$, $p < 0.001$ and $\chi^2(1) = 137.58$, $p < 0.001$, respectively, and a significant interaction, $\chi^2(2) = 9.72$, $p < 0.01$ and $\chi^2(2) = 16.86$, $p < 0.001$, respectively.

For the age and age group models, Tukey post hoc tests for coding scheme revealed that Exclude and Failure schemes marginally (diff = 0.12, SE = 0.054, CI = [0.01, 0.23], $z = 2.12$, $p < 0.10$) and significantly differed (diff = 0.14, SE = 0.054, CI = [0.03, 0.25], $z = 2.59$, $p < 0.05$) respectively. A similar analysis between Exclude and Ignore showed that they marginally differed for age (diff = 0.12, SE = 0.054, CI = [0.01, 0.23], $z = 2.10$, $p < 0.10$) and significantly differed for age group (diff = 0.14, SE = 0.054, CI = [0.03, 0.25], $z = 2.53$, $p < 0.05$). For Failure vs. Ignore, the difference was significant for both the age (diff = 0.23, SE = 0.050, CI = [0.13, 0.33], $z = 4.53$, $p < 0.001$) and age group models (diff = 0.28, SE = 0.050, CI = [0.18, 0.38], $z = 5.53$, $p < 0.001$).

We explored the effect of coding scheme further by analyzing the main effects and interactions within the mixed-effects models. In both the coding scheme × age and coding scheme × age-group mixed-effects models, the Exclude scheme was coded as the intercept and the main effects of the Failure scheme was significantly different from it, $t(2316) = -2.12$, SE = 0.054, $p < 0.05$, $r = 0.04$ and $t(2316) = -2.59$, $p < 0.01$, SE = 0.054, $r = 0.05$, respectively. The same was true looking at the difference between Exclude and Ignore, $t(2316) = 2.08$, SE = 0.054, $p < 0.05$, $r = 0.04$ and $t(2316) = 2.53$, SE = 0.054, $p < 0.05$, $r = 0.05$, respectively. As expected, participant’s age significantly affected their accuracy on the unexpected contents task, whether coded as age or age group, $t(1230) = 13.36$, SE = 0.004, $p < 0.001$, $r = 0.36$ and $t(1230) = 11.85$, SE = 0.050, $p < 0.001$, $r = 0.32$, respectively. There were marginal interactions between the Failure scheme and the age (or age group) factor, $t(2316) = 1.44$,

[6] Because this sample was gathered from a large set of individual experiments, there were other children for whom we attempted to administered the procedure, but not tested because they failed to answer this initial question. Unfortunately, it is not possible to recover the exact number of children who fit into this category, although we believe it is rare.

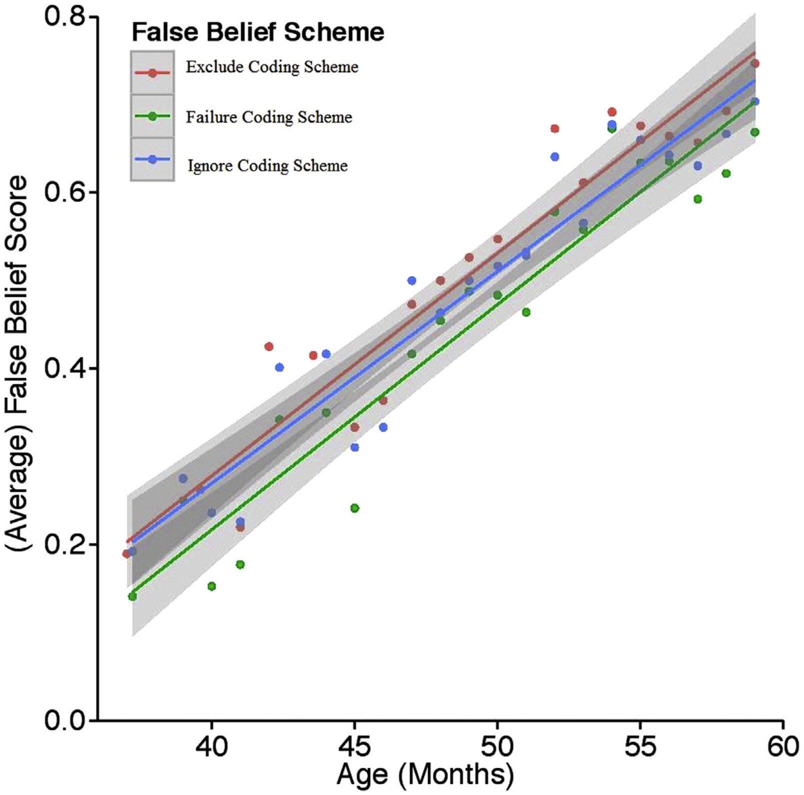

Fig. 3. Performance on false belief questions for the in-house data analysis depending on age (in months). For each question type, 10% quantiles are plotted as well as the best-fit logistic regression.

SE = 0.001, $p < 0.15$, $r = 0.03$ and $t(2316) = 1.91$, SE = 0.014, $p < 0.10$, $r = 0.04$, respectively, and the Ignore scheme and the age (or age group) factor, $t(2316) = 1.46$, SE = 0.001, $p < 0.15$, $r = 0.03$, and $t(2316) = 1.92$, SE = 0.014, $p < 0.10$, $r = 0.04$, respectively.

Overall, the coding scheme used can result in significantly different results. That is, this first set of analyses suggests that the coding schemes yield significant differences in overall accuracy scores among one another.

5.2.2. Simulated experiment analysis

So far, we have found that there is a reliable difference in the results of the unexpected contents task depending on the coding scheme used by the experimenter. Further, we replicated a result of the logit analysis – that the Exclude scheme differs significantly from the Failure and Ignore schemes. However, this analysis used a large number of children (over 1200) to establish the effect. Given the number of children typically used in developmental experiments, how strong of an effect does the coding scheme have on the results of the unexpected contents task?

We analyzed the effect size of coding scheme in a set of simulated experiments, focusing on the sample size of the hypothetical experiment and whether age is treated as a continuous (age) or nominal (age group) factor.[7] To perform a simulated experiment, we sampled (with replacement[8]) an equal number of 3-year-olds and 4-year-olds. We then analyzed the results of the simulated experiment using both the coding scheme x age and coding scheme × age group mixed models and stored the effect size of coding Schemes 2 and 3 (relative to coding Scheme 1), the age variable (for comparison) and the interaction between coding scheme and the age variable (for both models). We used $N = 30$, 40, 60, 100, and 200 total participants in the simulated experiments (so each experiment had $N/2$ three-year-olds and $N/2$ four-year-olds) and stored the results of analyzing the mixed models for 10,000 simulations of each sample size. Fig. 4 plots the distribution of effect sizes r (Pearson's correlation coefficient) for using Exclude, Failure, and Ignore with respect to each other in these simulated experiments, using either the *coding scheme* × *age* or *coding scheme* × *age group* model.

Typically, the effect sizes are small, but non-negligible, and approximately equal regardless of the coding scheme or model used. For example, when there are 30 simulated participants, the median effect size was greater than 0.09 and only

[7] Technically, age group should be treated as an ordinal variable – however, we find few papers that analyzed age group in this manner (i.e., via ordinal modeling). Most analyses treat age group as a nominal variable (as in ANOVA).

[8] The results are approximately the same if participants are sampled without replacement.

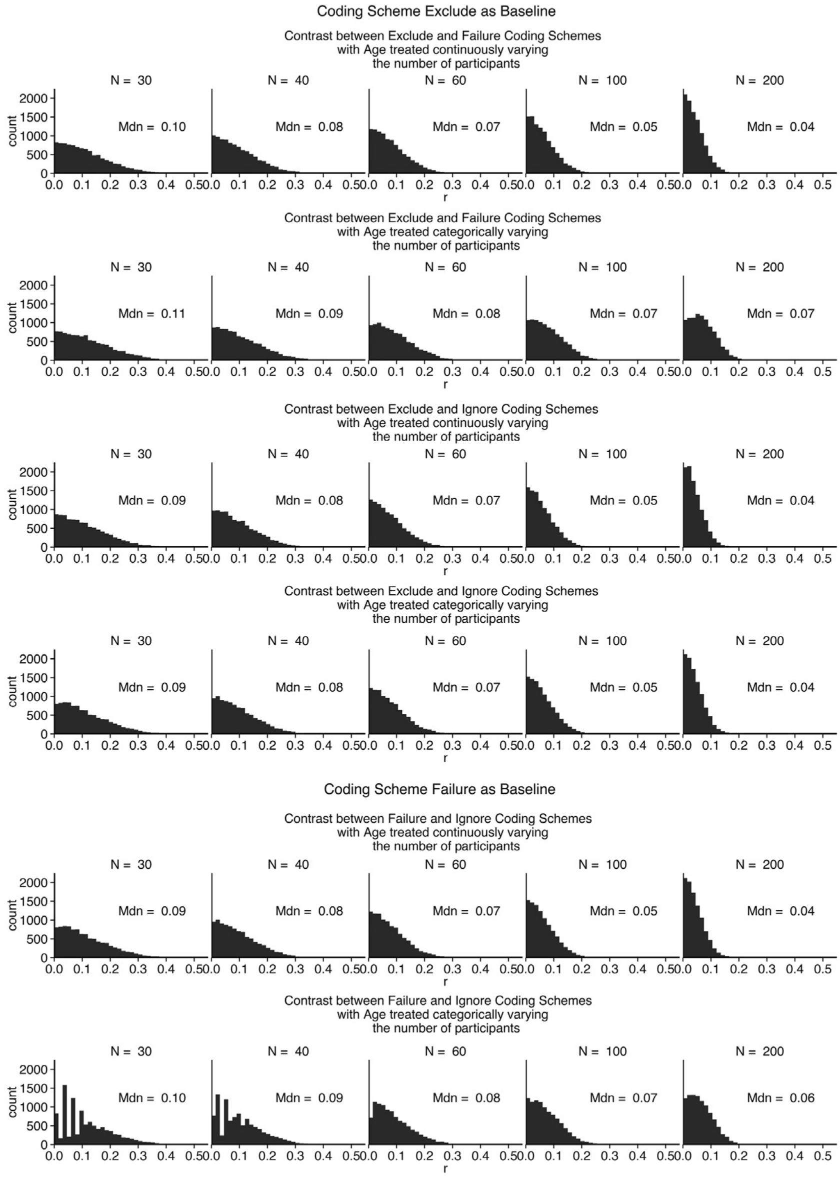

Fig. 4. The effect size (r) of comparisons among each of the three coding schemes in simulated experiments varying the number of participants and whether age was treated as a continuous covariate (through an ANCOVA, denoted *age* in the text) or in which age was treated categorically (via a coding scheme × age group ANOVA, denoted *age group* in the text). Mdn is the median effect size for each distribution. CI is the 95% confidence interval for the effect size. All pairs of coding schemes were tested. The effect size of between pairs of coding schemes is approximately equal for simulated experiments with typical number of participants (30–60) and decreases with increasing number of participants (regardless of the model used to analyze the results). The effect size of using the Failure coding scheme as opposed to the Exclude or Ignore systems decreases at a slower rate when the coding scheme × age group model is used, but not when the coding scheme × age model is used.

differed by 0.02 between the different models. Note this is about two times larger than the effect sizes when all participants are included in the analysis. Thus, the coding scheme has a stronger effect when the number of participants typical of most experiments is used, regardless of the model used for analysis.

As the effect size differences are relatively small with large sample sizes, it does provide some comfort that using different coding schemes has only a small effect on experimental results, especially when age is treated as a continuous variable. But to make these analyses more concrete, the median sample size from our literature review was ~60. Thus, these analyses suggest that at least one quarter of the reviewed literature showed effects with age that could be skewed by a change in Pearson Correlation of approximately 0.07 or greater (with the most noticeable difference between switching between the Exclude and Failure scheme). Given that even at N = 200 we are not using the entire sample, this is a conservative estimate and the actual effects are likely higher. Finally, the strongest difference here is between the Exclude and Failure schemes, which is consistent with the results of the logit analysis – in which the strongest effect of coding system was between Exclude and the other schemes, and no difference between Failure and Ignore.

6. Best practices and general discussion

We have shown that there is no consensus in how the unexpected contents measure is scored, and that the most frequently used scoring system based on our review did not factor in any control measure into analyses of the task. Further, across several analyses, we found that the coding scheme researchers used to score the unexpected content tasks potentially contributes to the significance level of analyses reported by the researchers. As such, we suspect that the coding scheme used for scoring this measure (and potentially related measures of false belief) could be a "researcher degree of freedom" (Simmons, Nelson, & Simonsohn, 2011, p. 1559) present in studies of children's theory of mind. We do not suspect that malice or deceptive intentions underlie any researcher when they choose how to score this measure. Indeed, we suspect that some authors have used a particular coding scheme for long periods, and this presumably reflects a genuine belief that there is an unambiguous way of scoring the measure. But this seems to reflect a minority; the majority of researchers using this measure appear to succumb to the genuine ambiguity in how to best score this measure.

We want to point out a particular limitation with our review, which is that we only focused on the unexpected contents task. Other measures of false belief (e.g., unexpected transfer, appearance-reality, etc.) also use control questions or pretests to ensure that children have the requisite attentional or memory capacities necessary to demonstrate their theory of mind knowledge. We limited our analysis to the unexpected contents task for two reasons. The first was practical – we had available to us a data set on children's performance on this particular measure. The second was that in the unexpected contents procedure, there tends to be only one control question asked (about the actual contents of the box), which simplified the review of the literature and our categorization of the different coding schemes. To our knowledge, other measures of false belief have not been analyzed in the manner presented here. We suspect that this analysis is representative of the broader literature on explicit judgments about false belief.

We want to recommend several best practices specific for using the unexpected content measure. We do not think these are the only best practices for developmental researchers generally. For instance, we encourage the use of a priori power analyses to justify sample sizes and the reporting all effect sizes and confidence intervals where appropriate. One way of addressing many concerns here is to pre-register studies or analyses, but frank and open discussion of the choices made during analyses can also address many issues brought up here.

First, we endorse the use of the Exclude coding scheme in almost all circumstances. Across all of our analyses, this scheme appears to provide the most consistent scoring of the measure. We can think of two exceptions to this endorsement. The first is when eliminating data is highly costly because the data comes from hard-to-collect populations. The second is when researchers are only interested in whether children have all the necessary capacities to make an inference about a false belief, including the memory capacity to remember those beliefs. The Failure scheme does ensure that children who pass the measure understand the task, even if overall, it is less valid than the Exclude scheme. Please note that both of these conditions were in place for the Peterson et al. (2005) data that we analyzed; we agree with their analyses and conclusions. We see no reason to ever use the Ignore scheme.

Our second best practice is to reconsider findings that exist in the literature for whether the coding scheme used by researchers matters. Of the papers we reviewed, 36 asked a control question, but used either the Failure or Ignore schemes. An open question is whether the findings of these papers (and presumably many others we did not consider) replicate when the Exclude scheme is used. Authors might consider making this information (in a de-identified manner, of course) available.

Our third best practice is for there to be a deeper discussion of the justification for the particular coding scheme, not just for this task, but for the many measures used in psychology. Consistent with the first practice we recommend above, the use of the Failure scheme should be the exception rather than the rule – and papers that present it should have detailed discussion of its usage as opposed to the Exclude scheme, including the trade-off between power and validity.

Fourth, regarding the treatment of age as a continuous or categorical variable, we recommend treating age as continuous given that it will provide a more accurate representation of the factor, provide more statistical power, and make analyses more robust to differences in coding scheme. If researchers insist on treating age as a categorical variable, they should consider subtracting the difference in effect size between Age Group and Age when doing a priori power analyses for sample sizes (see Fig. 4).

Our final best practices are for reviewers and editors. Do not accept manuscripts that use the Ignore scheme. Instead, insist that authors who use this coding scheme reanalyze their data. Control questions should be asked and factored into how this measure is scored. Moreover, we hope that developmental journals in particular are more receptive to papers reporting these kinds of methodological and statistical issues. Given the prevalence of new statistical techniques for data analysis, it seems necessary for the field to discuss how to make choices regarding scoring systems that can affect analyses.

To conclude, we have presented a case study in the literature on children's false belief, which we believe is representative of other theory of mind measures. Broadly speaking, it is worth discussing in any paper how raw data are scored. Based on the desire to achieve publishable findings, we believe researchers can score measures in a way that achieves significant results, and then engage in post-hoc justifications regarding why the measures were scored in that way. For common measures used across the field, like the unexpected contents task, this seems unacceptable. We hope that this paper inspires other researchers to think carefully about the scoring systems used for other measures, both in theory of mind and elsewhere in developmental science.

We also want to emphasize that our purpose in presenting this analysis is not to point out the flaws or methods with any particular paper or theory within the field of theory of mind. Rather, given the ubiquity of false belief measures in developmental science, we want to highlight that a best practice for scientific investigation in social cognitive development is to justify why data are scored in a certain way, and increase the robustness of known findings. We hope that the discussions brought up in this article generalize to other aspects of developmental science and other fields of psychology so that we can strive towards a better science.

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953.

Appleton, M., & Reddy, V. (1996). Teaching three year-olds to pass false belief tests: A conversational approach. *Social Development*, *5*(3), 275–291.

Astington, J. W., & Jenkins, J. M. (1999). A longitudinal study of the relation between language and theory-of-mind development. *Developmental Psychology*, *35*(5), 1311–1320.

Atance, C., & O'Neill, D. (2004). Acting and planning on the basis of a false belief: Its effects on 3-year-old children's reasoning about their own false beliefs. *Developmental Psychology*, *40*(6), 953–964.

Atance, C., Bernstein, D., & Meltzoff, A. (2010). Thinking about false belief: It's not just what children say but how they say it. *Cognition*, *116*(2), 297–301.

Baron-Cohen, S. (1991). Do people with autism understand what causes emotion? *Child Development*, *62*(2), 385–395.

Bartsch, K., & Wellman, H. (1989). Young children's attribution of action to beliefs and desires. *Child Development*, *60*(4), 946–964.

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.

Bernstein, D. M., Atance, C., Meltzoff, A. N., & Loftus, G. R. (2007). Hindsight bias and developing theories of mind. *Child Development*, *78*(4), 1374–1394.

Bialystok, E., & Senman, L. (2004). Executive processes in appearance–reality tasks: The role of inhibition of attention and symbolic representation. *Child Development*, *75*(2), 562–579.

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*(2), 647–663.

Blair, C., Granger, D., & Razza, R. P. (2005). Cortisol reactivity is positively related to executive function in preschool children attending head start. *Child Development*, *76*(3), 554–567.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, *72*(4), 1032–1053.

Carlson, S. M., Mandell, D. J., & Williams, L. (2004). Executive function and theory of mind: Stability and prediction from ages 2–3. *Developmental Psychology*, *40*(6), 1105–1122.

Carlson, S. M., Moses, L. J., & Claxton, L. J. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. *Journal of Experimental Child Psychology*, *87*(4), 299–319.

Cassidy, K. W., Werner, R. S., Rourke, M., Zubernis, L. S., & Balaraman, G. (2003). The relationship between psychological understanding and positive social behaviors. *Social Development*, *12*(2), 198–221.

Cassidy, K. W., Fineberg, D. S., Brown, K., & Perkins, A. (2005). Theory of mind may be contagious, but you don't catch it from your twin. *Child Development*, *76*(1), 97–106.

Cheung, H., Hsuan-Chih, C., Creed, N., Ng, L., Ping Wang, S., & Mo, L. (2004). Relative roles of general and complementation language in theory-of-mind development: Evidence from Cantonese and English. *Child Development*, *75*(4), 1155–1170.

Courtin, C. (2000). The impact of sign language on the cognitive development of deaf children the case of theories of mind. *Journal of Deaf Studies and Deaf Education*, *5*(3), 266–276.

Dalke, D. E. (1995). Explaining young children's difficulty on the false belief task: Representational deficits or context-sensitive knowledge? *British Journal of Developmental Psychology*, *13*(3), 209–222.

De Villiers, J. G., & Pyers, J. E. (2002). Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development*, *17*(1), 1037–1060.

Farrar, M. J., & Maag, L. (2002). Early language development and the emergence of a theory of mind. *First Language*, *22*(2), 197–213.

Fisher, N., Happé, F., & Dunn, J. (2005). The relationship between vocabulary, grammar, and false belief task performance in children with autistic spectrum disorders and children with moderate learning difficulties. *Journal of Child Psychology and Psychiatry*, *46*(4), 409–419.

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology*, *50*(1), 21–45.

Flynn, E., O'Malley, C., & Wood, D. (2004). A longitudinal, microgenetic study of the emergence of false belief understanding and inhibition skills. *Developmental Science*, *7*(1), 103–115.

Flynn, E. (2006). A microgenetic investigation of stability and continuity in theory of mind development. *British Journal of Developmental Psychology*, *24*(3), 631–654.

Freeman, N. H., & Lacohée, H. (1995). Making explicit 3-year-olds' implicit competence with their own false beliefs. *Cognition*, *56*(1), 31–60.

Frye, D., Zelazo, P. D., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development*, *10*(4), 483–527.

Geren, J., Snedeker, J., & Shafto, C. (2009). *The link between language and theory of mind: evidence from internationally adopted children*. Harvard University. Unpublished manuscript.

Glass, G. V., McGraw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage Publishing.

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development*, 26–37.

Guajardo, N. R., & Turley-Ames, K. J. (2004). Preschoolers' generation of different types of counterfactual statements and theory of mind understanding. *Cognitive Development*, *19*(1), 53–80.

Guajardo, N. R., & Watson, A. C. (2002). Narrative discourse and theory of mind development. *The Journal of Genetic Psychology*, *163*(3), 305–325.

Hala, S., Hug, S., & Henderson, A. (2003). Executive function and false-belief understanding in preschool children: Two tasks are harder than one. *Journal of Cognition and Development*, *4*(3), 275–298.

Hale, C. M., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: A training study. *Developmental Science*, *6*(3), 346–359.

Hogrefe, G. J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A developmental lag in attribution of epistemic states. *Child Development*, *57*(3), 567–582.

Howell, D. C. (2006). *Statistical Methods for Psychology*. Belmont, CA: Wadsworth Publishing.

Hughes, C., & Cutting, A. L. (1999). Nature, nurture, and individual differences in early understanding of mind. *Psychological Science*, *10*(5), 429–432.

Hughes, C., Jaffee, S. R., Happé, F., Taylor, A., Caspi, A., & Moffitt, T. E. (2005). Origins of individual differences in theory of mind: From nature to nurture? *Child Development*, *76*(2), 356–370.

Hughes, C. (1998). Executive function in preschoolers: Links with theory of mind and verbal ability. *British Journal of Developmental Psychology*, *16*(2), 233–253.

Jackson, A. L. (2001). Language facility and theory of mind development in deaf children. *Journal of Deaf Studies and Deaf Education*, *6*(3), 161–176.

Jenkins, J. M., & Astington, J. W. (1996). Cognitive factors and family structure associated with theory of mind development in young children. *Developmental Psychology*, *32*(1), 70–78.

Kalish, C. W., Weissman, M. D., & Bernstein, D. (2000). Taking decisions seriously: Young children's understanding of conventional truth. *Child Development*, *71*(5), 1289–1308.

Kelley, E., Paul, J. J., Fein, D., & Naigles, L. R. (2006). Residual language deficits in optimal outcome children with a history of autism. *Journal of Autism and Developmental Disorders*, *36*(6), 807–828.

Krachun, C., Carpenter, M., Call, J., & Tomasello, M. (2009). A competitive nonverbal false belief task for children and apes. *Developmental Science*, *12*(4), 521–535.

Lackner, C. L., Bowman, L. C., & Sabbagh, M. A. (2010). Dopaminergic functioning and preschoolers' theory of mind. *Neuropsychologia*, *48*(6), 1767–1774.

Lackner, C., Sabbagh, M. A., Hallinan, E., Liu, X., & Holden, J. J. (2012). Dopamine receptor D4 gene variation predicts preschoolers' developing theory of mind. *Developmental Science*, *15*(2), 272–280.

Lalonde, C. E., & Chandler, M. J. (1995). False belief understanding goes to school: On the social-emotional consequences of coming early or late to a first theory of mind. *Cognition & Emotion*, *9*(2–3), 167–185.

Leslie, A. M., & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition*, *43*(3), 225–251.

Lewis, C., & Osborne, A. (1990). Three-year-olds' problems with false belief: Conceptual deficit or linguistic artifact? *Child Development*, *61*(5), 1514–1519.

Lind, S. E., & Bowler, D. M. (2009). Recognition memory, self-other source memory, and theory-of-minf in children with Autism Spectrum Disorder. *Journal of Abnormal Psychology*, *39*, 1231-–1239.

Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. *Child Development*, *74*(4), 1130–1144.

Low, J., Goddard, E., & Melser, J. (2009). Generativity and imagination in autism spectrum disorder: Evidence from individual differences in children's impossible entity drawings. *British Journal of Developmental Psychology*, *27*(2), 425–444.

Lundy, J. E. (2002). Age and language skills of deaf children in relation to theory of mind development. *Journal of Deaf Studies and Deaf Education*, *7*(1), 41–56.

Müller, U., Zelazo, P. D., & Imrisek, S. (2005). Executive function and children's understanding of false belief: How specific is the relation? *Cognitive Development*, *20*(2), 173–189.

Müller, U., Miller, M. R., Michalczyk, K., & Karapinka, A. (2007). False belief understanding: The influence of person, grammatical mood, counterfactual reasoning and working memory. *British Journal of Developmental Psychology*, *25*(4), 615–632.

Major, A., Franco, F., & Zotović, M. (2010). Theory of mind and preschoolers' understanding of implicit causality in verbs: A comparison between Serbian and Hungarian children. *Psihologija*, *43*(2), 187–198.

Mathews, R., Dissanayake, C., & Pratt, C. (2003). The relationship between theory of mind and conservation abilities in children using an active/inactive paradigm. *Australian Journal of Psychology*, *55*(1), 35–42.

Mitchell, P., & Lacohée, H. (1991). Children's early understanding of false belief. *Cognition*, *39*(2), 107–127.

Moore, C., Pure, K., & Furrow, D. (1990). Children's understanding of the modal expression of speaker certainty and uncertainty and its relation to the development of a representational theory of mind. *Child Development*, *61*(3), 722–730.

Moses, L. J., & Flavell, J. H. (1990). Inferring false beliefs from actions and reactions. *Child Development*, *61*(4), 929–945.

Naito, M., Komatsu, S. I., & Fuke, T. (1994). Normal and autistic children's understanding of their own and others' false belief: A study from Japan. *British Journal of Developmental Psychology*, *12*(3), 403–416.

Pellicano, E., Maybery, M., & Durkin, K. (2005). Central coherence in typically developing preschoolers: Does it cohere and does it relate to mindreading and executive control? *Journal of Child Psychology and Psychiatry*, *46*(5), 533–547.

Pellicano, E. (2010). Individual differences in executive function and central coherence predict developmental changes in theory of mind in autism. *Developmental Psychology*, *46*(2), 530–544.

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, *5*(2), 125–137.

Perner, J., Frith, U., Leslie, A. M., & Leekam, S. R. (1989). Exploration of the autistic child's theory of mind: Knowledge, belief, and communication. *Child Development*, *60*(3), 689–700.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Peterson, C. C., & Siegal, M. (1999). Representing inner worlds: Theory of mind in autistic, deaf, and normal hearing children. *Psychological Science*, *10*(2), 126–129.

Peterson, C. C., Wellman, H. M., & Liu, D. (2005). Steps in theory-of-mind development for children with deafness or autism. *Child Development*, *76*(2), 502–517.

Peterson, C. C. (2000). Kindred spirits: Influences of siblings' perspectives on theory of mind. *Cognitive Development*, *15*(4), 435–455.

Repacholi, B., & Trapolini, T. (2004). Attachment and preschool children's understanding of maternal versus non-maternal psychological states. *British Journal of Developmental Psychology*, *22*(3), 395–415.

Robinson, E. J., Riggs, K. J., & Samuel, J. (1996). Children's memory for drawings based on a false belief. *Developmental Psychology*, *32*(6), 1056–1064.

Ruffman, T., Olson, D. R., Ash, T., & Keenan, T. (1993). The ABCs of deception: Do young children understand deception in the same way as adults? *Developmental Psychology*, *29*(1), 74–87.

Ruffman, T., Slade, L., Rowlandson, K., Rumsey, C., & Garnham, A. (2003). How language relates to belief, desire, and emotion understanding. *Cognitive Development*, *18*(2), 139–158.

Sabbagh, M. A., Bowman, L. C., Evraire, L. E., & Ito, J. (2009). Neurodevelopmental correlates of theory of mind in preschool children. *Child Development*, *80*(4), 1147–1162.

Saltmarsh, R., Mitchell, P., & Robinson, E. (1995). Realism and children's early grasp of mental representation: Belief-based judgements in the state change task. *Cognition*, *57*(3), 297–325.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Slade, L., & Ruffman, T. (2005). How language does (and does not) relate to theory of mind: A longitudinal study of syntax, semantics, working memory and false belief. *British Journal of Developmental Psychology*, *23*(1), 117–141.

Slaughter, V., & Gopnik, A. (1996). Conceptual coherence in the child's theory of mind: Training children to understand belief. *Child Development*, *67*(6), 2967–2988.

Slaughter, V., Dennis, M. J., & Pritchard, M. (2002). Theory of mind and peer acceptance in preschool children. *British Journal of Developmental Psychology*, *20*(4), 545–564.

Smith, M., Apperly, I., & White, V. (2003). False belief reasoning and the acquisition of relative clause sentences. *Child Development*, *74*(6), 1709–1719.

Sobel, D. M. (2004). Children's developing knowledge of the relation between mental awareness and pretense. *Child Development*, *75*, 704–729.

Sobel, D. M. (2006). How fantasy benefits young children's understanding of pretense. *Developmental Science*, *9*, 63–75.

Sobel, D. M. (2015). Can you do it?: How preschoolers judge others have learned. *Journal of Cognition and Development*, *16*, 492–508.

Sobel, D. M., Sommerville, J. A., Travers, L. V., Blumenthal, E. J., & Stoddard, E. (2009). Preschoolers' use of others' beliefs to make causal inferences from probabilistic data. *Journal of Cognition and Development*, *10*, 262–284.

Symons, D. K., Peterson, C. C., Slaughter, V., Roche, J., & Doyle, E. (2005). Theory of mind and mental state discourse during book reading and story-telling tasks. *British Journal of Developmental Psychology*, *23*(1), 81–102.

Tardif, T., Wellman, H. M., & Cheung, M. (2004). False belief understanding in Cantonese-speaking children. *Journal of Child Language*, *31*(04), 779–800.

Taylor, M., & Carlson, S. M. (1997). The relation between individual differences in fantasy and theory of mind. *Child Development*, *68*(3), 436–455.

Taylor, M., Lussier, G. L., & Maring, B. L. (2003). The distinction between lying and pretending. *Journal of Cognition and Development*, *4*(3), 299–323.

Tine, M., & Lucariello, J. (2012). Unique theory of mind differentiation in children with autism and Asperger syndrome. *Autism Research and Treatment*, *2012*.

Van Reet, J., Green, K. F., & Sobel, D. M. (2015). Preschoolers' existing theory of mind knowledge influences whom they trust about others' theory of mind. *Journal of Cognition and Development*, *16*, 471–491.

Vinden, P. G. (1996). Junin Quechua children's understanding of mind. *Child Development*, *67*(4), 1707–1716.

Walker, S. (2005). Gender differences in the relationship between young children's peer-related social competence and individual differences in theory of mind. *The Journal of Genetic Psychology*, *166*(3), 297–312.

Watson, A. C., Nixon, C. L., Wilson, A., & Capage, L. (1999). Social interaction skills and theory of mind in young children. *Developmental Psychology*, *35*(2), 386–391.

Watson, A. C., Painter, K. M., & Bornstein, M. H. (2001). Longitudinal relations between 2-year-olds' language and 4-year-olds' theory of mind. *Journal of Cognition and Development*, *2*(4), 449–457.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523–541.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684.

Williams, D. M., & Happé, F. (2009). What did I say? Versus what did I think? Attributing false beliefs to self amongst children with and without autism. *Journal of Autism and Developmental Disorders*, *39*(6), 865–873.

Wimmer, H., & Hartl, M. (1991). The Cartesian view and the theory view of mind: Developmental evidence from understanding false belief in self and other. *British Journal of Developmental Psychology*, *9*(125), 125–138.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, *1*(2), 217–235.

Zelazo, D., Jacques, S., Burack, J. A., & Frye, D. (2002). The relation between theory of mind and rule use: Evidence from persons with autism-spectrum disorders. *Infant and Child Development*, *11*(2), 171–195.