Replication of 'The Curse of Knowledge in Reasoning about False Beliefs'
by Susan A.J. Birch & Paul Bloom (2007, *Psychological Science*)

Sophie Bridgers
Michael C. Frank

Stanford University
Department of Psychology
450 Serra Mall, Jordan Hall, Building 420
Stanford, California 94305
(sbridge@stanford.edu)

Introduction

We spend a great deal of time reasoning about other people's thoughts and beliefs in order to better explain their past actions and predict their future actions. Young children struggle with this type of perspective taking, and in particular, have difficulty reasoning about others' false beliefs -- or beliefs that conflict with reality. Birch and Bloom (2007) attempt to better understand the source of this difficulty. Some researchers have claimed that children find false belief reasoning challenging because of conceptual limitations, such as lacking a concept of a belief or a mental representation at all (e.g., Gopnik, 1993; Perner, Leekam, & Wimmer, 1987; Wellman, 1990; Wellman, Cross, & Watson, 2001). While other researchers have proposed that children's difficulties are due to cognitive limitations, such as lower memory and processing capacities (e.g., Fodor, 1992; German & Leslie, 2000; Leslie, 1987; Onishi & Baillargeon, 2005; Roth & Leslie, 1998; Zaitchik, 1990). Birch and Bloom provide evidence in support of this latter view.

Birch and Bloom argue that the *curse of knowledge* bias, difficulty ignoring your own knowledge when reasoning about a less informed perspective, that is present in both children and adults, may be more pronounced in childhood and could explain children's deficits in false belief reasoning. The study shows that when sensitive enough measures are used, even adults' existing knowledge can interfere with their ability to reason about false beliefs. To demonstrate this, they used a modified displacement task. In the standard task, an agent (the protagonist) leaves an item in a particular container. Another agent then moves the item to a different container, while the protagonist is away, and the question is where the protagonist will look for the item when she returns. In their modified task, Birch and Bloom manipulated whether the participants knew where the item was relocated (the *Knowledge* conditions vs. the *Ignorance* condition) and whether there was a plausible explanation for why the protagonist would look in the container where the item actually was rather than in the container where she left it (the *Knowledge-plausible* condition vs. the *Knowledge-implausible* condition). Results reveal that when adults know the specific outcome of the displacement, they find it more probable that the protagonist will look in the true-belief location than when they are ignorant of the outcome, and that the influence of this knowledge bias is mediated by the plausibility of the outcome.

Here, we aim in particular to replicate two findings:
1. Adults in the *Knowledge-plausible* condition provided significantly higher probabilities that the protagonist would look in the true-belief container and significantly lower probabilities that the protagonist would look in the false-belief container than in the *Ignorance* condition (true-belief: $t(105) = -2.42$, $p_{rep} = 0.95$, $d = 0.472$; false-belief: $t(105) = 2.35$, $p_{rep} = 0.95$, $d = 0.459$), and

2. Adults' probability judgments did not significantly differ between the *Knowledge-implausible* and *Ignorance* conditions (true-belief: $t(97) = -1.44$, $p_{rep} = 0.95$, *n.s.*; false-belief: -0.21, $p_{rep} = 0.95$, *n.s.*).

Methods

Power Analysis

For the true-belief finding described above, the original effect size was $d = 0.472$. Power analyses revealed that to detect this effect size with 80%, 90%, and 95% power, the sample size would need to be 71 participants per condition, 95 participants per condition, or 118 participants per condition, respectively.

For the false-belief finding described above, the original effect size was $d = 0.459$. Power analyses revealed that to detect this effect size with 80%, 90%, and 95% power, the sample size would need to be 75 participants per condition, 101 participants per condition, or 124 participants per condition, respectively. These sample sizes will be feasible given that the replication will be conducted on Amazon Mechanical Turk.

All power analyses were conducted using `power.t.test()` in R with delta = {0.472 or 0.459}, sd = 1, sig.level = 0.05, power = {.80, .90, or .95}, type = "two.sample".

Planned Sample

Two hundred twenty-five adults will be tested via an online survey on Amazon Mechanical Turk. Participants will be randomly assigned to one of three conditions: the *Ignorance* condition (n = 75), the *Knowledge-plausible* condition (n = 75), and the *Knowledge-implausible* condition (n = 75). Participants will be paid $0.20 for their time. Participation will be restricted to Amazon Mechanical Turk workers who have a HIT acceptance rate of 85% and U.S. IP addresses. We will aim to have balanced numbers of male and female participants, but female participants may be over-represented in our sample due to the fact that approximately 70% of U.S. Amazon Mechanical Turk workers are female.

Materials and Procedure

The following Materials and Procedure were followed precisely and were identical to those used in Birch and Bloom (2007), with the following critical exceptions: 1. the survey was an electronic

survey administered online rather than a paper survey administered in person; 2. the entire survey could not be viewed at once (i.e., participants needed to scroll down to see all of the images and text); 3. the images were re-created to be a very close approximation to those used in the original experiment, but they were not identical; the main differences being the shape of the purple container, that none of the boxes have patterns, and that Denise is playing the violin rather than holding it; 4. important information was bolded; 5. color-words were included next to the corresponding response box. Please see Figure 1 for images of the original stimuli and the stimuli used in this replication.

“All subjects received the same stimuli...The first picture depicted a girl who was holding a violin and standing by a sofa and four containers. Each container was a different color: blue, purple, red, and green. Beneath the first picture was an image of a different girl holding a violin; in this picture, the same four containers were rearranged.

Subjects in all conditions read, ‘This is Vicki. She finishes playing her violin and puts it in the blue container. Then she goes outside to play. While Vicki is outside playing, her sister, Denise . . . .’ At this point, the conditions differed:

Ignorance: ‘moves the violin to another container.’
Knowledge-plausible: ‘moves the violin to the red container.’
Knowledge-implausible: ‘moves the violin to the purple container.’

All subjects then read, ‘Then, Denise rearranges the containers in the room until the room looks like the picture below.’ This was followed by, ‘When Vicki returns, she wants to play her violin. What are the chances Vicki will first look for her violin in each of the above containers?’” (Birch & Bloom, 2007, pp 383). Type your answers in percentages in the spaces provided under each container.

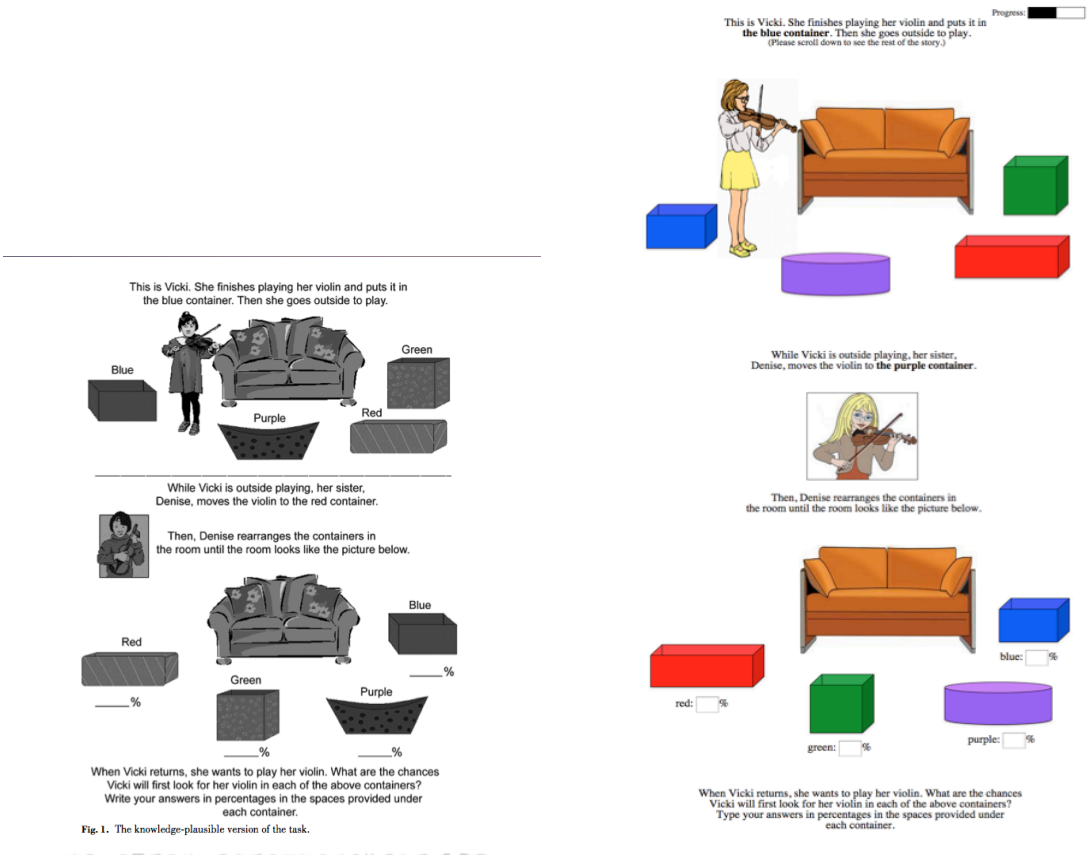


Figure 1. Black and white version of original stimuli for the *Knowledge-plausible* condition on the left (the actual survey used was in color and did not include color words). Replication stimuli for the *Knowledge-implausible* condition on the right.

Analysis Plan

No exclusion criteria were described in the original study. We will exclude participants based on the following criteria: 1. If a participant takes the survey more than once, only their first set of responses will be included in analysis; 2. Participants who fail to answer any of the questions will be excluded from analysis.

We will replicate the two-sample *t*-tests conducted in the original study to see if there are differences in participants' probability judgments of how likely the protagonist (Vicki) will be to look in the true-belief and false-belief containers across the three conditions (i.e., we will compare each condition to the other two conditions). We will also fit the data with a mixed-effects model with condition and container as fixed effects and a random effect of intercept for subjects.

Differences from Original Study

The planned sample size (N = 225) is larger than the sample size of the original study (N = 155). This increased sample size will allow us to detect the observed effect in the original study with 80% power. The original study sample consisted only of Yale undergraduates. The composition of the planned sample will be more diverse in age and background. A critical difference of the planned materials and procedure is that the survey will be administered electronically online rather than on paper in person. This difference, plus the other differences outlined above are not anticipated to make a difference in the findings based on claims in the original article or subsequent research.

(Post Data Collection) Methods Addendum

Actual Sample

Two hundred twenty-five participants were recruited via Amazon Mechanical Turk and were compensated $0.20 for their time. Participants were randomly assigned to one of three conditions: the *Ignorance* condition (n = 65), the *Knowledge-implausible* condition (n = 73), and the *Knowledge-plausible* condition (n = 87). All participants had U.S. IP addresses and an 85% HIT approval rating. Participants were 34% female and 30.31 years old on average (sd = 8.99).

Differences from pre-data collection methods plan

Due to the random assignment of the participants to conditions, there were slightly fewer participants than planned in the *Ignorance* and *Knowledge-implausible* conditions and slightly more in the *Knowledge-plausible* condition. There were no other differences from the pre-data collection methods plan.

Results

Data preparation

No participants were excluded from analysis. We normalized participants' probability judgments for how likely the protagonist (Vicki) would be to look in each of the four containers. For the planned *t*-test comparisons, we subset the data by condition.

In order to fit the data with a mixed model, we gathered by container and response to transform the data from wide to long form. Additionally, due to the co-linearity of participants' normalized responses (they summed to 100), we removed participants' probability judgments for the green container since no predictions were made about this container across conditions.

Confirmatory analysis

We aimed to replicate two major findings from the original study: 1. Adults in the

Knowledge-plausible condition provide significantly higher probabilities that the protagonist would look in the true-belief container and significantly lower probabilities that the agent would look in the false-belief container than in the *Ignorance* condition, and 2. Adults' probability judgments do not significantly differ between the *Knowledge-implausible* and *Ignorance* conditions.

To test finding (1), we conducted two-sample *t*-tests comparing participants' probability judgments for the blue (false-belief) container and red (true-belief) container in the *Ignorance* and *Knowledge-plausible* conditions. We did not replicate this result: We found *no* significant differences in participants' responses for these containers between the two conditions (blue container: $t(128.55) = -0.52$, $p = 0.604$; red container: $t(149.58) = -1.20$, $p = 0.231$). When participants knew the violin was in the red container, they were no less likely to say that Vicki would look in the blue container nor were they more likely to say she would look in the red container as compared to when they did *not* know where the violin had been moved.

To test finding (2), we again conducted two-sample *t*-tests but this time comparing participants' probability judgments for the blue (false-belief) container and purple (true-belief) container in the *Ignorance* and *Knowledge-implausible* conditions. We did replicate this result: There were no significant differences in participants' responses for these containers between the two conditions (blue container: $t(128.67) = -1.27$, $p = 0.206$; purple container: $t(133.89) = 0.36$, $p = 0.716$). When participants knew the violin was in the purple container, they were no less likely to say that Vicki would look in the blue container nor were they more likely to say she would look in the purple container than when they did *not* know to which container Denise had moved the violin.

The following analyses were part of our analysis plan but were not reported in the original paper.

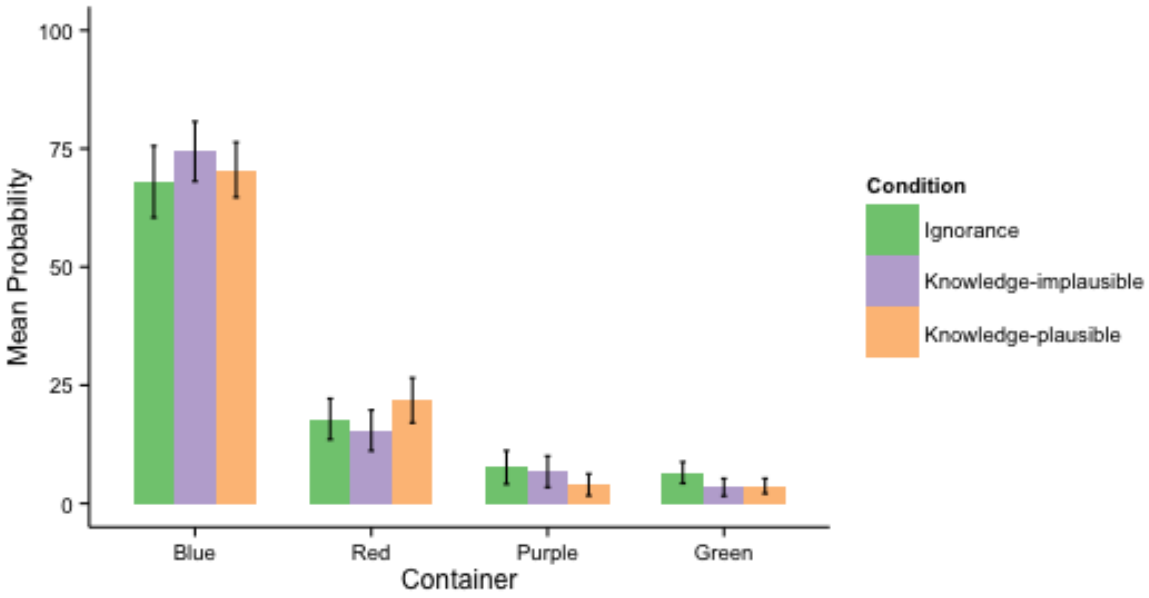
We conducted two-sample *t*-tests comparing participants' probability judgments in the *Knowledge-plausible and Knowledge-implausible* conditions for the false-belief container (blue in both conditions) and the true-belief container (red in the *Knowledge-plausible* condition and purple in the *Knowledge-implausible* condition). We found no difference in participants' responses for the false belief (blue) container between these two conditions ($t(153.52) = -0.88$, $p = 0.379$). We did find a significant difference in participants' responses for the true belief (red/purple) container across conditions ($t(147.94) = 5.106$, $p < 0.001$). Knowing the violin was in the red or the purple container did not lead to a significant difference in how likely participants' thought Vicki would be to act in accordance with a false-belief. However, participants thought Vicki was much more likely to act in accordance with a true-belief if the violin was in the red container than if it was in the purple container.

Lastly, we fit the data with an interactive mixed effects model with condition and container as fixed effects and a random effect of intercept for subjects (see equation 1 for R code). We used orthogonal contrasts to compare participants' probability judgments for the blue vs. the red containers in the *Ignorance* and *Knowledge-plausible* conditions (i.e., finding 1 from the original study) and to compare participants' probability judgments for the blue vs. the purple containers

in the *Ignorance* and *Knowledge-implausible* conditions (i.e., finding 2 from the original study). The model revealed that participants judged it more likely that Vicki would look in the blue (false-belief) container than in the red or the purple containers across all three conditions (blue - red: b = 25.071, t = 13.22; blue - purple: b = 30.18, t = 15.92). Consistent with the results of the t-tests, the model additionally revealed that there was no difference in participants' probability judgments of how likely Vicki would be to look in the blue container vs. the red container in the *Ignorance* condition compared to the *Knowledge-plausible* condition (b = -0.702, t = -0.28). Also consistent with the results of the t-tests, the model revealed no difference in participants' probability judgments of how likely Vicki would be to look in the blue vs. the purple container in the *Ignorance* condition compared to the *Knowledge-implausible* condition (b = 3.650, t = 1.40). Results from the original study and from this replication attempt are summarized in Figure 2.

```
lmer(response_norm ~ cond * container + (1 | workerid), d.noGreen)
```
(1)

(a)



(b)

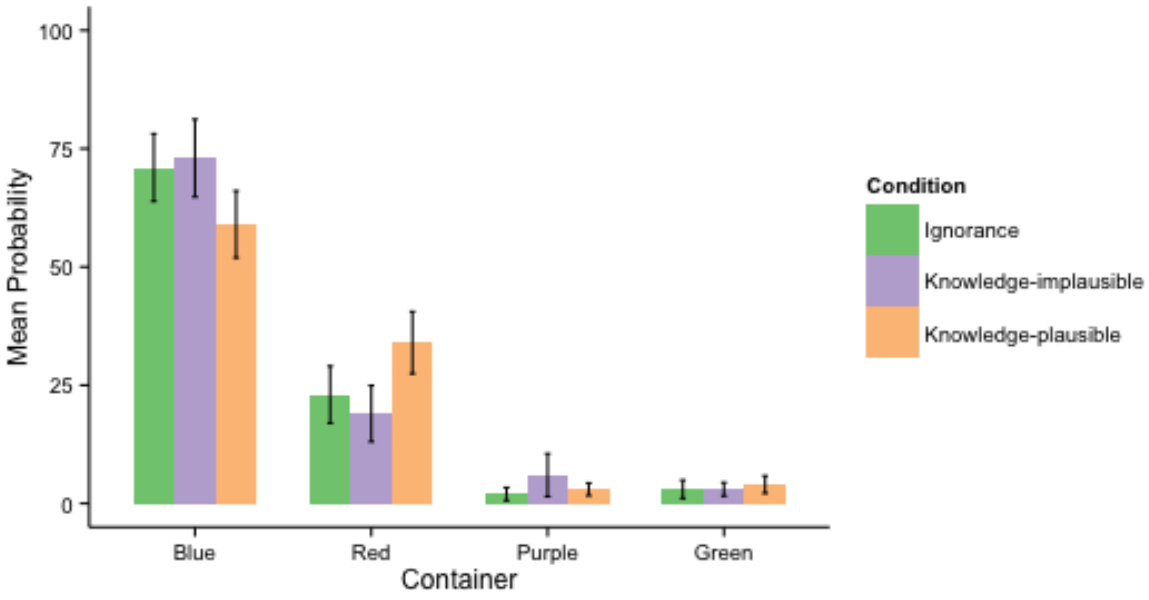


Figure 2. Graph (a) summarizes the data from the replication attempt and Graph (b) summarizes the data from the original study. The graphs plot participants' average judgments of how likely Vicki would be to look in each of the containers for each of the conditions. Error bars represent 95% confidence intervals.

Exploratory analyses

We ran two additional two-sample *t*-tests comparing participants' probability judgments for 1. the red container and 2. the purple container in the *Knowledge-plausible* and *Knowledge-implausible* conditions. The *t*-tests revealed a trending significant difference between these two conditions in participants' responses for the red container ($t(157.99) = 1.94$, $p = 0.054$) but no significant difference for the purple container ($t(133.15) = -1.34$, $p = 0.182$). Knowing the violin was in the red container marginally increased the plausibility for participants that Vicki would first look in the red container compared to knowing the violin was in the purple container. However, knowing the violin was in the purple container did not similarly increase the plausibility for participants that Vicki would first look there compared to knowing the violin was in the red container.

Discussion

Summary of Replication Attempt

We found that there were no significant differences in participants' probability judgments of how likely the protagonist would be to first look in the true-belief container or the false-belief container in the *Knowledge-plausible* and the *Ignorance* conditions. We also found no

differences when comparing participants' probability judgments for these containers in the *Knowledge-implausible* and *Ignorance* conditions. In other words, we found that participants' knowledge of where the violin had been moved did *not* affect their judgments of how likely the protagonist would be to first look in the true-belief container (i.e., where the violin was actually located). We did not replicate the original result that knowing the violin was located in the red container increased participants' judgments of the plausibility Vicki would first look there and decreased the plausibility that she would first look in the blue container compared to when they had no knowledge of the violin's location. We did, however, replicate the original result that knowing the violin was in the purple container did *not* affect participants' probability judgments compared to when they had no knowledge of where the violin had been moved. In sum, we partially replicated the results from the original study.

Birch and Bloom (2007) concluded that having knowledge of where the violin is actually located interferes with adults' ability to take Vicki's perspective but only when there is also a plausible reason why Vicki would act in accordance with their own knowledge rather than with her own. However, we found that participants' knowledge of where the violin had been moved did *not* affect their judgments of how likely the protagonist would be to act in accordance with a false-belief, regardless of the plausibility of the true-belief.

The plausibility manipulation in our replication appears valid. Participants in the *Ignorance* condition gave the red container a mean probability rating of 18% compared to an 8% rating for the purple container and a 7% rating for the green container. These ratings indicate that when participants had no knowledge of the violin's true location, they thought it was relatively plausible Vicki would first look for the violin in the red container (the initial location of the blue container) and relatively implausible she would first look in the purple container. Additionally, participants thought it was significantly more likely Vicki would act in accordance with a true-belief when the violin was in the red container than when it was in the purple container, further supporting that where the blue container was originally located is a probable place for Vicki to look relative to other locations. However, the partial failure of our replication suggests that participants in our study did not exhibit the 'curse of knowledge' bias found in the original study.

Commentary

Though our confirmatory analyses failed to demonstrate the 'curse of knowledge' bias in action, our exploratory analyses provide some evidence in favor of it. These analyses revealed that participants in the *Knowledge-plausible* condition predicted Vicki would be more likely to look in the red container (where the violin was actually located) than in the *Knowledge-implausible* condition when the violin was in the purple container. This finding, though marginal, suggests that participants' knowledge of where the violin was located may have biased their judgments, increasing their assessment of how likely it would be for Vicki to look in that container (e.g., the red container) compared to when they knew the violin was in a different container (e.g., the purple container).

Our power analyses prior to data collection indicated that having 75 participants would allow us to detect the observed effect in the original study with 80% power. However, due to random assignment of participants to condition, we were slightly under-powered in the *Ignorance* condition which may have prevented us from detecting a difference between this condition and the *Knowledge-plausible* condition. It is also possible that using different stimuli and the fact that the entire survey could not be viewed at once (i.e., participants needed to scroll down to see all of the images and text) impeded our ability to fully replicate the original study. Participants in the *Ignorance* condition gave the red container a mean probability rating of 18% compared to an 8% rating for the purple container, suggesting they did think it was relatively plausible Vicki would first look for the violin in the red container and relatively implausible she would first look in the purple. Thus, the plausibility manipulation was in effect, however, to a lesser extent than in the original study where participants in the *Ignorance* condition gave a mean probability rating of 23% to the red container and 2% to the purple container. Future replication attempts should address these concerns.

References

Birch, S. A., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*(5), 382-386.Fodor, J. (1992). A theory of the child's theory of mind. Cognition, 44, 283–296.

German, T.P., & Leslie, A.M. (2000). Attending to and learning about mental states. In P. Mitchell & K. Riggs (Eds.), Children's reasoning and the mind (pp. 229–252). Hove, England: Psychology Press.

Gopnik, A. (1993). How we know our own minds: The illusion of first person knowledge of intentionality. Behavioral and Brain Sciences, 16, 1–14.

Leslie, A.M. (1987). Pretense and representation: The origins of ‘‘theory of mind.’’ Psychological Review, 94, 412–426.

Onishi, K.H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? Science, 308, 255–258.

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, 125–137.

Roth, D., & Leslie, A.M. (1998). Solving belief problems: Towards a task analysis. Cognition, 66, 1–31.

Wellman, H.M. (1990). The child's theory of mind. Cambridge, MA: MIT Press.

Wellman, H.M., Cross, D., & Watson, J. (2001). Meta-analysis of theory of mind development:

The truth about false-belief. Child Development, 72, 655–684.

Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and ''false'' photographs. Cognition, 35, 41–69.