Infant Behavior and Development 54 (2019) 177–188

Contents lists available at ScienceDirect

Infant Behavior and Development

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

Full length article

Publication standards in infancy research: Three ways to make Violation-of-Expectation studies more reliable

Paula Rubio-Fernández[a,b]

[a] Massachusetts Institute of Technology, Department of Brain & Cognitive Sciences, 43 Vassar St., Building 46, Cambridge, MA, 02139, USA
[b] University of Oslo, Department of Philosophy, Georg Morgenstiernes Hus, Blindernveien 31, Oslo, 0313, Norway

ARTICLE INFO

Keywords:
Surprise
False-belief reasoning
End-of-trial criteria
p-curve analysis
Piloting
Accuracy
Eye-tracking measures

ABSTRACT

The Violation-of-Expectation paradigm is a widespread paradigm in infancy research that relies on looking time as an index of surprise. This methodological review aims to increase the reliability of future VoE studies by proposing to standardize reporting practices in this literature. I review 15 VoE studies on false-belief reasoning, which used a variety of experimental parameters. An analysis of the distribution of p-values across experiments suggests an absence of p-hacking. However, there are potential concerns with the accuracy of their measures of infants' attention, as well as with the lack of a consensus on the parameters that should be used to set up VoE studies. I propose that (i) future VoE studies ought to report not only looking times (as a measure of attention) but also looking-away times (as an equally important measure of distraction); (ii) VoE studies must offer theoretical justification for the parameters they use, and (iii) when parameters are selected through piloting, pilot data must be reported in order to understand how parameters were selected. Future VoE studies ought to maximize the accuracy of their measures of infants' attention since the reliability of their results and the validity of their conclusions both depend on the accuracy of their measures.

1. Introduction

In a climate of general scepticism towards psychological research (Simmons, Nelson, & Simonsohn, 2011; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012; Yong, 2012), Peterson (2016) recently published an ethnographic study in which he described questionable practices that were frequently observed in three infant cognition labs. Peterson's study focused on data collection with the so called 'Violation-of-Expectation paradigm' (henceforth VoE), which is an extensively used paradigm in infancy research. The VoE relies on eye-tracking measures of total looking time that are normally recorded by a human observer and interpreted as an index of surprise. In Onishi and Baillargeon (2005), for example, 15-month-old infants looked at a scene in which an agent was mistaken about the location of an object and reached for it either in the container where it used to be before (expected outcome) or where it was now (unexpected outcome). Infants looked longer at the final scene in the unexpected condition, which was interpreted as evidence that they understood that the agent had a false belief about the location of the object and were surprised that she did not act accordingly.

This methodological review also focuses on the VoE paradigm. However, rather than targeting questionable data-collection practices in the lab, in the spirit of Simmons et al. (2011), I will focus on the methodological reports of these studies with the aim of raising the publication standards of the field and standardizing experimental parameters in the long run. The VoE paradigm has been used for over two decades to investigate research questions as diverse as infants' understanding of addition and subtraction (Wynn,

E-mail address: prubio@mit.edu.

https://doi.org/10.1016/j.infbeh.2018.09.009
Received 14 May 2018; Received in revised form 8 August 2018; Accepted 28 September 2018
Available online 26 October 2018

1992), goal-directed action (Csibra, Bíró, Koós, & Gergely, 2003; Gergely, Nádasdy, Csibra, & Biró, 1995; Király, Jovanovic, Prinz, Aschersleben, & Gergely, 2003; Luo & Baillargeon, 2005; Woodward, 1998), the role of perception in action (Luo & Baillargeon, 2007; Luo & Johnson, 2009) and efficient action (Scott & Baillargeon, 2013). Therefore, the methodological issues discussed in this review are broadly relevant to developmental psychology and infant cognition research. However, in order to illustrate critical methodological issues, the present review will focus on recent VoE studies on psychological reasoning, while its conclusions apply more generally to infant studies using the VoE paradigm.

A corpus of 15 VoE studies published to date on false-belief reasoning in infancy (He, Bolz, & Baillargeon, 2011; Kovács, Téglás, & Endress, 2010; Luo, 2011; Onishi & Baillargeon, 2005; Scott, 2017; Scott & Baillargeon, 2009; Scott, Baillargeon, Song, & Leslie, 2010; Scott, He, Baillargeon, & Cummins, 2012; Scott, Richman, & Baillargeon, 2015; Song & Baillargeon, 2008; Song, Onishi, Baillargeon, & Fisher, 2008; Surian, Caldi, & Sperber, 2007; Träuble, Marinović, & Pauen, 2010; Yott & Poulin-Dubois, 2012), including a cross-cultural replication (Barrett et al., 2013), will be used to illustrate the methodological limitations of the VoE paradigm. There is an ongoing debate around this literature as to whether infants have a 'Theory of Mind' that allows them to pass these tasks by attributing false beliefs to agents (Butterfill & Apperly, 2013; Falck, Brinck, & Lindgren, 2014; Fenici, 2014; Heyes, 2014; Perner & Ruffman, 2005; Rakoczy, 2015; Ruffman, 2014; Wellman, 2014). Recently, Powell, Hobbs, Bardis, Carey, and Saxe (2017) published a failed replication of the original study by Onishi and Baillargeon (2005), which adds to a number of failed replications of early Theory of Mind studies (cf. Baillargeon et al., 2018; Grosse Wiesmann, Friederici, Singer, & Steinbeis, 2017; Kammermeier & Paulus, 2017; Kulke, Reiß, Krist, & Rakoczy, 2017; Kulke, von Duhn, Schneider, & Rakoczy, 2018; Rubio-Fernández, 2018; Yott & Poulin-Dubois, 2016). However, while the interpretation of the original results has been controversial, and the reliability of the findings seems to be currently under question, the methodological limitations of the VoE paradigm have not been so carefully examined in the context of either debate (for the only methodological review to date, see Sirois & Jackson, 2007). Therefore, rather than contributing to the interpretation and replication debates, the present review will focus instead on the methodological limitations of the VoE paradigm.

The review will be divided in two parts. First, given the variability of the experimental parameters used in most false-belief VoE studies to date, an analysis of the distribution of p-values across experiments (*p-curve analysis*; Simonsohn, Nelson, & Simmons, 2014a, 2014b) was conducted in order to investigate whether these parameters may have been selected post-hoc to optimize p-value. Importantly, the main analysis suggests that the p-curve is 'healthy', supporting the studies' evidential value. However, this does not itself validate the experimental parameters used in those VoE studies. In particular, the wide variability in parameter settings leaves two important questions unanswered:

(1) Have all these variable parameters maximised (or even preserved) the accuracy of their measures of infants' attention?
(2) How should future VoE studies set their experimental parameters in the absence of any objective norms for parameter setting?

These two questions will be addressed in the second part of the paper. There I will focus on three issues concerning the reliability of the results of VoE studies and argue that (a) measures of both attention and distraction need to be reported in order to accurately estimate infants' expectations, (b) the selection of experimental parameters needs to be theoretically motivated, and (c) in those instances where pilot data are used to select experimental parameters, the pilot data need to be reported in order to ensure that the selection was not simply based on infants' performance. I conclude by arguing that future VoE studies should also employ more advanced eye-tracking techniques that allow combining various looking measures, rather than relying on a single measure of overall looking time.

2. End-of-trial criteria

In the VoE paradigm, infants' attention is often measured from the onset of a paused scene following a series of events that are acted out for the infant. The end of a trial is normally controlled by the infant – what is called an *infant-controlled procedure*. This method was introduced in the 1980s and Woodward (1998) used the following criteria for determining when an infant has lost interest in a paused scene: the trial ends when (a) the infant looks away for 2 consecutive seconds or (b) 120 s have elapsed without the infant looking away for 2 consecutive seconds. After these thresholds, it is assumed that the infant is no longer surprised by the outcome of the events. This twofold criterion is considered a standard convention in the field (although the 120-second maximum looking time has often been shortened; e.g., Hespos, Saylor, & Grossman, 2009; Kovács et al., 2010; Yott & Poulin-Dubois, 2012).

False-belief studies often refer to Woodward's work to motivate their methodology (Onishi & Baillargeon, 2005; Scott & Baillargeon, 2009; Scott et al., 2010, 2015; Song & Baillargeon, 2008; Song et al., 2008; Surian et al., 2007). However, of the 15 VoE studies published to date in that field, only 3 used a twofold criterion to end their trials (see Table 1), with the remaining 12 studies using other criteria (see Table 2 and Barrett et al., 2013). The latter studies will be referred to as *VoE studies with variable parameters* and will be the focus of the present review.

Most VoE studies with variable parameters used a combination of three looking measures: a disrupted-looking measure (i.e. a maximum look-away time) and a continuous-looking measure (i.e. a maximum looking time) that were used to end the trial, plus a cumulative minimum looking time, before which the trial could not end. These three looking measures are tiered: only after achieving the cumulative minimum looking time is it possible for the infant to end the trial by looking away for the maximum consecutive time. Similarly, only if the infant has not looked away for the maximum consecutive time is it possible for the infant to end the trial by reaching the maximum cumulative looking time.

In VoE studies, trials normally have an initial phase (when the action takes place) and a final phase (when the scene is paused),

Table 1
False-belief studies using the VoE paradigm with standard parameters. Test trials are highlighted in bold.

Article	Mean age (months)	Exp.	Trial Sequence	End of trial A Disrupted looking Maximum look-away time (consecutive sec)	End of trial B Continuous looking Maximum looking time (cumulative sec)
Surian et al. (2007)	13	1-2	5 familiarization **1 test**	**2**	**120**
Kovács et al. (2010)	7	4-7	2 familiarization **2 test**	**2**	**30**
Yott and Poulin-Dubois (2012)	18	1	3 familiarization 1 belief-induction **1 test**	**2**	**30**

Table 2
False-belief studies using the VoE paradigm with variable parameters. Test trials are highlighted in bold.

Article	Mean age (months)	Exp.	Trial Sequence	End of trial A Disrupted looking		End of trial B Continuous looking
				Minimum looking (cumulative seconds)	Maximum look-away (consecutive seconds)	Maximum looking time (cumulative/ *consecutive sec)
Onishi and Baillargeon (2005)	15	1	3 familiarization 1 belief-induction **1 test**	**2**	**2**	**30**
Song and Baillargeon (2008)	14.5	1	4 familiarization 1 box-orientation **1 test**	5 5 **6**	2 2 **2**	60 30 **30**
Song et al. (2008)	18	1-2	3 familiarization 1 belief-induction 1 intervention **1 test**	2 3 (Fixed duration) **5**	2 2 **2**	60 40 **60**
Scott and Baillargeon (2009)	17	1-3	4 familiarization **2 test**	5 **4**	2 **2**	60 **40**
Scott et al. (2010)	18	1-3	1 [Exp1, Exp3] / 2 [Exp2] familiarization **1 test**	15 **5**	2 **0.5**	60 **30**
Träuble et al. (2010))	15	1	3 familiarization 1 belief-induction **1 test**	**n.a.**	**n.a.**	**20***
He et al. (2011)	31	1	2 familiarization **1 test**	5 **9**	2 **1.5**	30 **60**
		2	1 familiarization **2 test**	5 **7**	2 **0.5**	30 **45**
Luo (2011)	10	1-2	2 orientation 3 familiarization **4 test**	2 2 **5**	2 2 **2**	10 30 **60**
Scott et al. (2012)	30	2	1 familiarization **1 test**	(Fixed duration) **12**	**0.5**	**60**
Scott et al. (2015)	17	1-2	6 familiarization **1 test**	5 **5**	2 **1**	60 **30**
		3	6 familiarization **1 test**	5 **5**	2 **1.5**	60 **30**
Scott (2017)	20	1	2 familiarization **1 test**	10 **15**	2 **2**	60 **50**
		2	2 familiarization **1 test**	10 **5**	2 **2**	60 **50**
		3	2 familiarization **1 test**	5 **5**	2 **2**	60 **50**

Table 3
F and *t* statistics from planned comparisons of expected vs. unexpected conditions in false-belief (critical) and true-belief (control) trials of VoE studies using variable parameters. Experiments that predicted a null result were not included in the p-curves.

Article	Exp.	False-Belief condition	True-Belief condition
Onishi and Baillargeon (2005)	1	Individual *F* statistics not reported	Individual *F* statistics not reported
Song and Baillargeon (2008)	1 (doll)	$F(1,40) = 18.07$	$F(1,40) = 5.58$
	1 (skunk)	$F(1,40) = 6.43$	$F(1,40) = 6.69$
Song et al. (2008)	1	$F(1,24) = 11.00$	$F(1,24) = 5.12$
	2	$F(1,28) = 12.97$	$F(1,28) = 5.86$
Scott and Baillargeon (2009)	1	$F(1,26) = 13.17$	$F(1,26) = 9.03$
	2	$F(1,26) = 7.83$	No sig. diff. predicted
	3	$F(1,42) = 10.90$	No sig. diff. predicted
Scott et al. (2010)	1	$F(1,32) = 7.78$	$F(1,32) = 4.21$
	2	$F(1,28) = 4.65$	$F(1,28) = 7.22$
	3	No sig. diff. predicted	n.a.
Träuble et al. (2010)	1	$t(11) = -3.00$	$t(11) = 1.90$
He et al. (2011)	1	$F(1,48) = 7.96$	$F(1,48) = 4.93$
	2	$F(1,42) = 10.41$	$F(1,42) = 11.30$
Luo (2011)	1	$F(1,22) = 4.61$	No sig. diff. predicted
	2	No sig. diff. predicted	$F(1,10) = 11.67$
Scott et al. (2012)	2	$F(1,24) = 10.22$	$F(1,24) = 9.17$
Scott et al. (2015)	1	$F(1,32) = 11.73$	No sig. diff. predicted
	2	$F(1,32) = 5.21$	No sig. diff. predicted
	3	$F(1,32) = 9.75$	No sig. diff. predicted
Scott (2017)	1	$F(1, 24) = 5.43$	No sig. diff. predicted
	2	$F(1, 24) = 8.81$	$F(1, 24) = 7.68$
	3	$F(1, 24) = 4.31$	$F(1, 24) = 9.21$

and looking times are calculated separately for each phase. The duration of the initial phase is normally fixed and infants' looking times during that phase are used as an indicator of their paying attention to the events. The criteria shown in Tables 1 and 2 were used to end the final phase of trials, once the final scene had been paused and critical looking times started being recorded.[1]

3. P-curve analyses of VoE studies with variable parameters

Given the wide variability in the end-of-trial criteria used in the studies in Table 2 and the lack of theoretical justification for these precise parameters, a statistical analysis was carried out in order to investigate whether those parameters might have been set post-hoc to optimize p-value. For this purpose, the distribution of critical p-values in a set of studies can be used to test for evidence of p-hacking in those studies (Simonsohn et al., 2014a, 2014b).

It may at first seem that using different end-of-trial criteria in VoE studies could not be a form of p-hacking since these criteria are supposed to be set prior to the start of the recording. However, because the entire testing session is normally video recorded in parallel with the looking-time capture, it is possible to re-code the original videos using a different set of end-of-trial criteria. This is, for example, what He at al. (2011, Footnote 2) did in order to re-analyse the results of Experiment 1 using the end-of-trial criteria from Experiment 2 (Matthias Bolz, email communication). It therefore seems legitimate to do p-curve analyses of the studies in Table 2.

In psychology studies, the null hypothesis can be rejected if a critical p-value is less than .05. If a set of studies investigates a robust effect, the distribution of p-values will be right-skewed such that there will be more p-values between 0 and .01 than between .02 and .03, and more p-values between .02 and .03 than between .03 and .04, etc. On the other hand, if the underlying effect is not real, there will be a flat distribution of p-values between 0 and 1, which suggests a lack of evidential value in the data. Finally, if researchers are actively p-hacking (i.e. re-running different analyses until a significant result emerges), then the p-curve will be left skewed such that there will be more p-values between .04 and .05 than between .03 and .04.

In VoE studies, the null hypothesis is that infants do not distinguish between what would be expected and unexpected outcomes for an adult. In the case of false-belief studies, the null hypothesis is that infants do not differentiate what is expected and unexpected depending on an agent's knowledge state. Consider again the design of Onishi and Baillargeon (2005), in which an agent puts an object inside a container and either leaves the scene and misses the object moving to a second container (false-belief condition) or stays in the scene and watches the object move to a second container (true-belief control). When the agent is right about the location of the object, she should look for it in its current location – and it would be surprising if she looked for it in the alternative, empty location. On the other hand, when the agent is wrong about the location of the object, she should look for it in the empty location where it was before – and it would be surprising if she looked for it in its current location.

In order to decide which p-value to use in a p-curve analysis, it is necessary to identify the main statistical prediction of each

[1] In He et al. (2011), Scott et al. (2012; Experiment 2) and Scott (2017; Experiment 1), the final scene was not paused; instead, the last action was repeated in a loop until the trial ended.

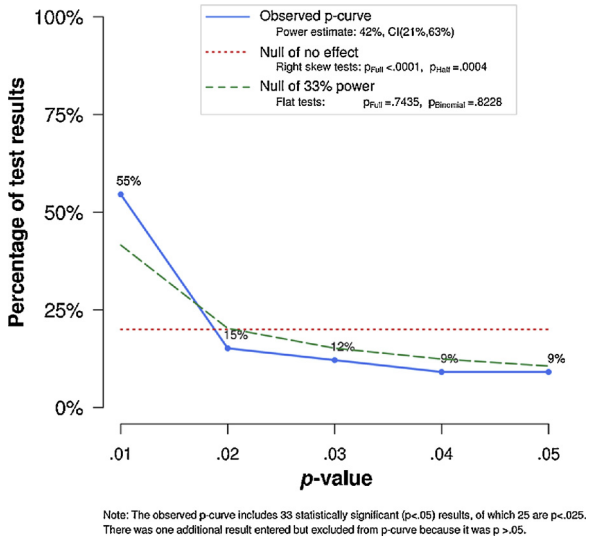

Fig. 1. Plot produced by the p-curve app 4.05 (http://www.p-curve.com/app4/) for false-belief VoE studies using variable parameters. The solid blue line is the distribution of p-values in the sample. The dotted red line shows the expected distribution of p values if there was no underlying effect. The dashed green line shows the expected p-curve under 33% statistical power. The right skew in the solid blue line shows there is evidential value in this set of studies (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.).

study. In the case of false-belief VoE studies, the main prediction is that infants will look longer at events that are unexpected than at events that are expected, showing sensitivity to whether the agent has a true or a false belief about a certain state of affairs. If an effect is predicted to reverse, the p-value of both simple effects should be included in the p-curve. Given that in true-belief and false-belief conditions opposite outcomes would be predicted, the *F* statistics of planned comparisons of expected vs. unexpected outcomes in each of the two conditions were entered in the analysis (see Table 3).

Following Simonsohn, Simmons, and Nelson (2015), p-curves are tested for significant right skew (i.e. most p-values near zero) using Stouffer's method. The observed p-curve is also compared to a hypothetical p-curve with 33% power since Simonsohn et al. (2015) argue that observed curves that are flatter than the 33% power curve suggest a lack of evidential value (for more details, see Simonsohn et al., 2014a, 2014b).

Fig. 1 plots the distribution of p-values for the studies in Table 3, including a total of 21 experiments and 34 tests. Most tests in the sample (55%) have p-values less than .01, and the curve shows significant right skew ($p < .0005$ by a Stouffer's z-test for skew on both the full p-curve and half p-curve; see Simonsohn et al., 2015). Comparing the solid blue line to the dashed green line shows that the number of small p-values is greater than one would expect if the power were 33%. The bias-corrected average power estimate is 42% with a 90% confidence interval of (21%, 63%).

The p-curve analysis gives no reason to suspect post-hoc manipulation, which is reassuring for the field of infant Theory of Mind. Note that this does not show that changing the end-of-trial criteria across experiments is innocuous: it merely reveals no evidence for directed malpractice.

Separate p-curve analyses were also carried out on the critical (false-belief) and control (true-belief) conditions. Whereas false-belief conditions enact situations where a character is mistaken about a certain state of affairs, the true-belief controls enact parallel situations where the character is correct about the same state of affairs. Therefore, what is unexpected in the critical condition (e.g., that the mistaken agent looks for the object in its current location) is expected in the control condition.

Fig. 2 plots the distribution of p-values for the false-belief conditions of the studies in Table 3, including a total of 20 tests. Most tests in the sample (65%) have p-values less than .01, and the curve shows significant right skew ($p < .0002$ by a Stouffer's z-test for skew on both the full p-curve and half p-curve). Comparing the solid blue line to the dashed green line shows that the number of small p-values is greater than one would expect if the power were 33%. The bias-corrected average power estimate is 52% with a 90% confidence interval of (25%, 75%).

Fig. 3 plots the distribution of p-values for the true-belief conditions, including a total of 14 tests. Only 38% of tests have p-values less than .01, and the curve does not show a significant right skew ($p < .05$ by a Stouffer's z-test for skew on the full p-curve and $p = .282$ on the half p-curve), therefore not indicating significant evidential value. The 33% power tests are not significant either, which indicates that evidential value is neither inadequate nor absent. The bias-corrected average power estimate is 26% with a 90% confidence interval of (5%, 62%).

The conclusion from these two p-curves is that whereas the false-belief curve reveals significant evidential value, the true-belief curve does not, even though neither p-curve is symptomatic of p-hacking. Theoretically speaking, however, there is no reason why infants should be more surprised if an agent acts contrary to a false belief than if she acts contrary to a true belief. In fact, the opposite might be expected.

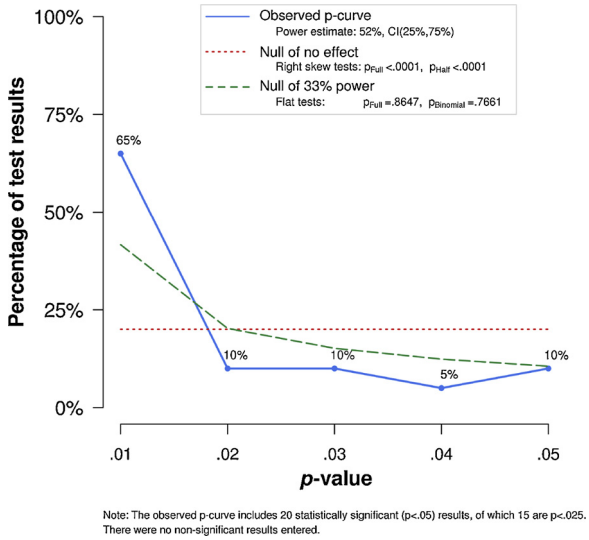

Fig. 2. Plot produced by the p-curve app 4.05 (http://www.p-curve.com/app4/) for the false-belief conditions of VoE studies using variable parameters. The right skew in the solid blue line shows there is evidential value in the false-belief conditions of these studies.

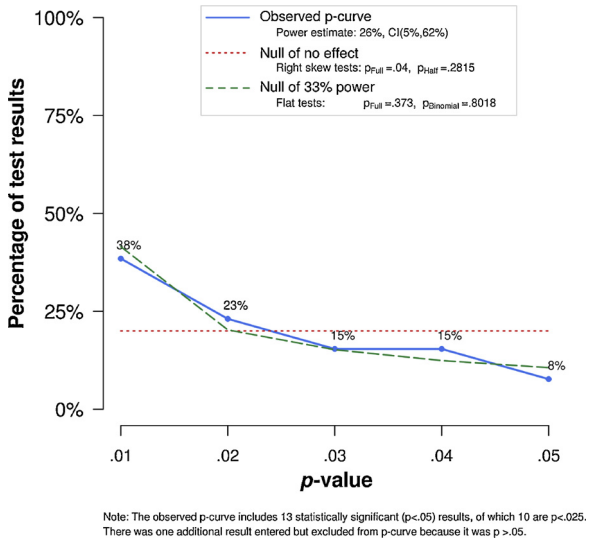

Fig. 3. Plot produced by the p-curve app 4.05 (http://www.p-curve.com/app4/) for the true-belief conditions of VoE studies using variable parameters. The relatively flat solid blue line shows there is no evidential value in the true-belief conditions of these studies.

Consider again the false-belief scenario in Onishi and Baillargeon (2005) in which a mistaken agent returns to a scene to look for an object: if she was to first reach in the container hiding the object, her behaviour may be initially surprising, but it would not be irrational. After all, the agent would have fulfilled her goal to obtain the object and it would be possible (at least for an adult) to reconstruct her behaviour rationally (e.g., she might have learned that the object was moved to the other container in her absence; see Song et al., 2008). On the other hand, if an agent is right about the location of an object and looks for it in a different, empty container, her behaviour does not satisfy her desire for the object, and neither can it be reconstructed as rational – why would anyone look for an object in a place different from where they know it to be?

Given that there is no theoretical reason why the index of surprise (i.e. the difference in looking times between expected and unexpected outcomes) should be greater in the false-belief conditions than in the true-belief conditions, this unexpected difference needs to be explained if the results of these VoE studies are to be taken at face value. One possibility consistent with this pattern of results is that the end-of-trial criteria that were selected for each experiment during piloting optimised infants' performance in the critical false-belief conditions, but not in the true-belief controls. Be that as it may, the different evidential value observed in critical and control conditions suggests that the end-of-trial parameters selected might not have accurately captured infants' attention, since doing so should have revealed surprise in all unexpected scenarios.

4. Methodological issues with VoE studies

Besides the results of p-curve analyses, the use of variable end-of-trial criteria without theoretical justification leaves two important questions unanswered. First, do all VoE studies with variable parameters accurately measure infants' attention? This is a key question because the use of variable end-of-trial criteria could allow the identification of significant differences in looking times between expected and unexpected outcomes, but these differences need not reveal surprise if the experimental parameters did not accurately capture infants' attention.

The second question is how can previous VoE studies inform future studies if they do not provide justification for not using standard parameters? This is clearly a pressing question for VoE researchers.

In what follows I will address these questions and propose three ways to increase the reliability of future VoE studies, which should in turn unify the experimental parameters used in this paradigm.

4.1. Issue 1: variable parameters and accuracy of measures

Aslin (2007) warned that caution must be exercised in what conclusions are drawn from looking time data since many unaccounted-for variables may contribute to this single measure. "The key to interpreting any given metric", Aslin argues, "is the linking hypothesis that joins the dependent variable to the underlying cognitive process" (p.49). The VoE paradigm measures overall looking time as an index of surprise, so if infants look longer at the final scene in Condition A than in Condition B, it is assumed that the events in Condition A were more surprising than those in Condition B. However, looking time is an indirect measure of surprise since it is possible both to look at a paused scene while thinking of something other than the previous events, and look away from the scene while still thinking about those events. This raises important issues about the reliability of VoE studies.

First, end-of-trial criteria are necessarily arbitrary and so there are no 'right' or 'wrong' criteria. Even the standard 2-second look-away criterion is a mere estimate of how long infants might be able to look away from a scene without getting distracted. Having said that, standardization has clear advantages in any scientific field, for it allows reliable comparison and replication of results across experiments. For example, the results of the false-belief VoE studies in Table 1 are more readily comparable than those in Table 2.

Moreover, the fact that there are no right or wrong end-of-trial criteria does not mean that some criteria are not better, or more reliable, than others. In particular, it is important that VoE studies use protocols that minimize the possibility of computing fixations on the scene that were made after an infant had lost interest in the events. This is precisely the purpose of the 2-second look-away criterion that is standardly used in VoE studies. However, most false-belief VoE studies did not conform to this standard: Träuble et al. (2010), for example, measured looking time during 20 consecutive seconds and allowed infants to look away from the scene for any length of time without ending the trial. In the remaining VoE studies using variable parameters (see Table 2), look-away criteria only applied after a cumulative minimum looking time had been achieved. This means that during a time window of 2– 15 cumulative seconds, depending on the experiment, infants were allowed to look away for any length of time without ending the trial.

VoE studies with test events that are repeated in a loop (instead of pausing the final scene) have sometimes used a minimum looking time to ensure that infants process the ongoing test events in sufficient detail (e.g., Gergely et al., 1995; Csibra et al., 2003). However, most false-belief VoE studies presented infants with a paused scene. In this paradigm, if infants look away for a few seconds and disengage from the scene before achieving the minimum looking time, then later fixations on the scene may be long because infants are trying to reconstruct the previous events, or are simply surprised that nothing is happening on stage – but not necessarily because the previous events had violated their expectations.

Whether infants look away from the scene before achieving a minimum looking time is likely to depend on the actual minimum established and the age of the infants. In critical/unexpected trials, the average looking times reported in the studies in Table 2 are relatively long: the longest being over 20 s in all of the studies, except Träuble et al. (2010) and Scott et al. (2015). This makes it likely that infants would have looked away from the scene during that time.

It must be noted that VoE studies from other labs have maximised the accuracy of their measures by applying both minimum looking times and the standard 2-second look-away criterion, plus exclusion criteria to deal with infants who did not achieve the minimum looking time (e.g., Gergely et al., 1995; Csibra et al., 2003; Király et al., 2003). It is therefore possible to ensure that infants have had enough time to process the events (which is the point of using a minimum looking time) while excluding those infants who may have got distracted in the meantime (which is the point of the standard 2-second look-away criterion).

4.2. Estimating the probability of computing 'disengaged fixations'

The longer the cumulative minimum looking time that infants must achieve before a trial can end, the greater the probability that infants will look away from the scene in that time. The false-belief studies in Table 2 used minimum looking times of 2, 4, 5 and 6 cumulative seconds with paused scenes in test trials. He et al. (2011); Scott et al. (2012; Experiment 2) and Scott (2017; Experiment 1) applied longer minimum looking times of 7, 9, 12 and 15 cumulative seconds using a repeated action in the final scene of test trials. In all these studies, infants' looking times were coded by human observers using basic software that did not record sequences of looks to the scene and away from the scene or their duration (for discussion, see Aslin, 2007). That means that the looking-time data are too coarse-grained to establish the proportion of infants who may have looked away from the scene for 2 s or more before achieving the cumulative minimum looking time.

This proportion is important, however, as it is key to assessing the reliability of the results of these studies: if a number of infants looked away from the scene for 2 consecutive seconds before achieving the cumulative minimum looking time, that means that a

Table 4
Proportions of infants in Surian et al. (2007) who looked away for 2 consecutive seconds (standard distraction criterion) before meeting the cumulative minimum looking times used in false-belief VoE studies using variable parameters with a paused final scene.

Surian et al. (2007)	Cumulative minimum looking time			
	2 sec	4 sec	5 sec	6 sec
Experiment 1 (N = 56)	.04	.09	.16	.20
Experiment 2 (N = 49)	.10	.23	.27	.31

certain proportion of the data should have been discarded according to the standard criterion for distraction, since fixations would no longer measure violation of expectation.

Unlike VoE studies with variable parameters, the eye-tracking data from VoE studies using standard parameters may be used to estimate the probability that infants may look away for longer than two seconds before reaching a minimum looking time (see Table 1). Thus, the looking-time data reported by Surian et al. (2007) shows that up to 27% of infants looked away from the scene for 2 s before achieving a 5-second minimum looking time, and up to 31% did so before achieving a 6-second minimum looking time (see Table 4).[2] Of course, these data are merely indicative of infants' looking behavior in this particular paradigm, and we do not know to what extent these results are replicated in the studies in Table 2.

It has been argued that using longer minimum looking times with paused scenes gives infants more time to process the events acted out in the initial phase of the trial and may therefore reveal better performance (Scott, 2017; Song & Baillargeon, 2008). However, the proportions in Table 4 suggest that using longer minimum looking times also undermines the reliability of the data as a larger proportion of infants may have looked away from the scene (hence potentially getting distracted) before achieving the minimum looking time.

To be able to assess the reliability of VoE studies using a cumulative minimum looking time, it is crucial that researchers report measures of looking-away time together with their looking times. Without knowing for how long infants looked away before achieving the minimum looking time, it is simply not possible to determine whether the reported looking times are a reliable index of surprise, or whether they are inflated by fixations on the scene that were made after infants had lost interest in the events. Therefore, a first proposal for increasing publication standards for VoE studies is to ensure that those studies using cumulative minimum looking times standardly report looking times and looking-away times as complementary measures of attention and distraction.

4.3. Issue 2: theoretical justification for variable parameters

The three looking measures used to end trials in most false-belief VoE studies were varied independently across conditions and studies, with a total of 18 different combinations of minimum looking time, maximum look-away time and maximum looking time (see Table 2). For test trials alone, these VoE studies used a total of 12 triple criteria in 10 papers.

Researchers often justify their choice of end-of-trial criteria in VoE studies (e.g., Gergely et al., 1995; Csibra et al., 2003; Király et al., 2003), but not always (e.g., Hespos et al., 2009). The motivation for the variable parameters used in the studies in Table 2 was only discussed in two of these papers. Since none of the other false-belief VoE studies offered any theoretical justification for their specific choice of experimental parameters, the ensuing discussion will focus on the explanations offered in these two papers:

> "Readers might wonder why a 5- or 6-s minimum looking time was required in the final phase of the familiarization, box-orientation, and test trials: because no event occurred during this phase (the infants simply watched a paused scene), why not use a short, 1- or 2-s, minimum looking time? In prior experiments on physical or psychological reasoning in which VOE tasks with paused scenes were used, we found that infants sometimes performed better with a slightly longer minimum looking time (e.g., 4 to 7 s), which gave them more time to process the information presented in the initial phase of the trial (e.g., Luo & Baillargeon, 2005, 2007; Song et al., 2008; Wang & Baillargeon, 2005). The specific minimum looking time used in each type of trial was established through piloting; here, because the test events differed from the previous events in several respects (e.g., the toys were fully hidden), a slightly longer minimum looking time seemed appropriate."
>
> Song & Baillargeon (2008:1792, Footnote 2)

> "Readers may wonder why different criteria were used to end the test trials in Experiments 1 and 2. The computer program we use to conduct VOE experiments (which is available free of charge on R. Baillargeon's website) allows investigators to set, for each research project, the parameters that best capture children's responses. These parameters are typically established during piloting and then used for data collection in the remainder of the project. Just as a whole host of factors can affect response parameters in visual-recognition tasks [References], many factors can affect response parameters in VOE tasks, including age, number of familiarization trials, similarity of the familiarization and test trials, complexity of the events shown in the test trials, and so on. Comparison of the criteria used to end trials in prior VOE false-belief tasks (Onishi & Baillargeon, 2005; Song & Baillargeon, 2008; Song et al., 2008; Surian et al., 2007; Träuble et al., 2010) reveals that each project had its own, slightly different set of criteria. Of

[2] I would like to thank Luca Surian and Dan Sperber for providing me with the looking-time data in Surian et al. (2007).

course, within any one project, conditions are run using the same criteria, so that conclusions are based on intra-project comparisons."

[The remainder of the footnote included a re-analysis of Experiment 1 using the end-of-trial criteria from Experiment 2 with a partial replication of the original results]

He et al. (2011:301, Footnote 2)

It is unclear from He et al.'s footnote precisely how their choice of end-of-trial criteria was informed by the various factors at play. Acknowledging that a measure of cognitive ability can be affected by a number of factors is hardly an argument for the use of variable experimental parameters without theoretical justification. After all, every measure of human ability can be affected by a number of factors, some of which will be related to the specifics of the task at hand. He et al. (2011) claim that their VoE studies use 'the parameters that best capture children's responses', but that does not explain how exactly those parameters are set. This does not mean that the parameters of these VoE studies must have been set arbitrarily, but it does call for more clarity and specificity when reporting their selection of end-of-trial criteria.

Given that data collection in VoE studies depends entirely on the specific end-of-trial criteria employed, the theoretical motivation for selecting one criterion rather than another ought to be explained in as much detail as the rest of the experimental procedure (which all VoE studies describe in great detail). Since the rationale for the selection of these parameters is not obvious from the age of the infants or the experimental procedure (see studies in Table 2), it should therefore be a publication requirement that researchers using the VoE paradigm with variable parameters justify their choice of end-of-trial criteria theoretically. Moreover, offering theoretical justification for criteria selection would not only make reported results more interpretable, but it would also have broad benefits for VoE research, as it should help standardization of looking parameters in the longer run.

4.4. Issue 3: using pilot data to select experimental parameters

Both papers discussed above argue that specific end-of-trial criteria are normally established through piloting. This explanation raises an important question for developmental research: to what extent is it valid to pilot the end-of-trial criteria used in a VoE study? On the one hand, one might argue that without piloting these criteria, infants' cognitive abilities might go undetected. On the other hand, it is also reasonable to argue that by piloting end-of-trial criteria (rather than consistently using standard criteria) the positive results of a VoE study might be artificially high – potentially resulting in the publication of false positives or idiosyncratic effects that do not reflect the intended ability. Thus, Gelman and Loken (2013, 2014) have recently shown how it is possible for researchers to construct reasonable rules for data exclusion, coding and analysis that may lead to statistical significance without having to perform multiple statistical analyses (i.e. without p-hacking).

Referring to previous studies, Song & Baillargeon (2008:1792, Footnote 2) also explained that 'infants sometimes performed better with a slightly longer minimum looking time'. Given the nature of the VoE paradigm, this explanation calls for closer examination: in what way could infants perform 'better' with a longer minimum looking time? Unlike standard false-belief tasks, which children may pass or fail (Wimmer & Perner, 1983), infants in false-belief VoE studies are only supposed to look at a paused scene for as long as they are interested. However, there are perhaps three ways in which infants might arguably perform better in VoE studies. First, infants must follow the events and pass the attention checks conducted in the initial phase of trials. However, these checks are not affected by the criteria used to end the final phase of trials. Second, infants should not become bored or fussy because the events are too long. This, again, does not depend on using a longer minimum looking time in the final phase. Third, infants might perform better by showing longer looking times in unexpected trials than in expected trials (as predicted). This aspect of infants' performance could indeed be directly affected by the choice of end-of-trial criteria.

A possible concern here is that an indicator of better performance need not reveal the most accurate results. It is possible that different end-of-trial criteria may reveal more accurate measures of infants' surprise, potentially at the expense of producing non-significant results. Therefore, it ought to be established that the findings of VoE studies that identified end-of-trial criteria through piloting are replicable using consistent standard criteria. Without replication or re-analysis, the end-of-trial criteria used in the VoE studies in Table 2 cannot be independently validated.

If future VoE studies continue to select end-of-trial criteria through piloting, then pilot data should be reported in order to ensure that these parameters were selected objectively. For example, in a VoE study on action and perception, Luo and Johnson (2009) described their second experiment as follows: 'The procedure was identical to that of Experiment 1 with one exception – based on pilot data, the maximum looking time in test trials was capped at 30 rather than 60 s' (p. 146). This description (and similar ones often found in VoE studies; e.g., Scott & Baillargeon, 2013) beg the following question: what exactly did the pilot data show to warrant that the maximum looking time should be halved?

The issue of what exactly pilot data may show in order to inform three different end-of-trial criteria is even more pressing when we consider the kind of looking-time data recorded in these studies. Whereas modern eye-trackers provide continuous looking-measures, the software used in the VoE studies in Table 2 generated very simple outputs including the duration of the look away that ended the trial and the looking time for that trial (Rose Scott, email communication; Yuyan Luo, email communication). This means that their pilot data did not provide information about the different looking patterns that may have emerged during the task, which could in turn inform the researchers' choice of end-of-trial criteria in the actual experiment.

One way in which coarse-grained pilot data may be informative with regards to the setting of parameters is by re-coding the pilot videos using different end-of-trial criteria and monitoring the results. This practice, however, would increase the possibility of reporting false positives by identifying highly specific parameters that do not preserve the accuracy of their measures of attention.

Therefore, the third way in which publication standards ought to improve for VoE studies is by having researchers report the pilot data that informed their choice of experimental parameters.

5. Reliable measures of surprise require reliable measures of attention

In view of the variable parameters used in most false-belief VoE studies, there are three reasons to suspect that the looking times reported in these studies may not reflect an accurate measure of infants' attention. First of all, the VoE studies in Table 2 used minimum looking times between 2– 15 cumulative seconds and Träuble et al. (2010) used a fixed maximum looking time of 20 consecutive seconds. Given the considerable length of these minimum looking times, infants may have looked away for 2 s and potentially got distracted before achieving their minimum. This is a serious cause for concern since looking time is supposed to indicate how surprising the previous events were, and distraction compromises the interpretation of ensuing fixations on the scene.

The second reason to suspect the reliability of the VoE studies in Table 2 is that the maximum look-away times were often shorter than 2 s, potentially stopping the recording before the infant got distracted. Scott et al. (2010, 2012, 2015) and He et al. (2011) used maximum look-away times of 1.5, 1.0 and 0.5 consecutive seconds without any evident theoretical justification. These studies used relatively long minimum looking times (5– 12 cumulative seconds), which begs the question of why infants would have got distracted from the events so quickly. In the case of the shortest minimum look-away time (0.5 s), looking times were probably stopped when the infant made a single fixation out, which may not necessarily indicate that the infant had already disengaged from the events.

These short look-away criteria are in clear contrast with other VoE studies in Table 2, which used similarly long minimum looking times but adopted the standard 2-second look-away criterion (Luo, 2011; Scott, 2017; Song & Baillargeon, 2008; Song et al., 2008). If all these studies are reporting accurate measures of infants' attention, it needs to be explained why after achieving comparable minimum looking times, some infants would have got distracted by looking away from the scene for 0.5 s, while other infants would still be attentive when looking away for 1.9 s.

The third reason to suspect that the VoE studies with variable parameters may not have accurately measured infants' attention is that when p-curve analyses are performed on the false-belief and true-belief conditions separately, the p-values from the false-belief conditions reveal evidential value, whereas the p-values from the true-belief conditions do not. Given that there is no reason why observing an agent acting against a false belief should be more surprising than watching an agent act against a true belief (in fact, the opposite may be true), the lack of evidential value of the control conditions casts doubts on the reliability of these results.

It must be noted that the accuracy of looking times as a measure of infants' attention is more fundamental to the conclusions of these and other VoE studies than whether or not the experimental predictions were borne out by the data. If their looking times do not accurately capture infants' attention to the events, then the fact that there was a significant difference between expected and unexpected outcomes cannot be taken to indicate surprise, let alone to support the theoretical claims of these studies.

In conclusion, the potential inaccuracy of looking times as a measure of infants' surprise raises questions about the interpretation of this body of work. These concerns can be alleviated by (i) reporting measures of both attention and distraction, in order to provide a more accurate estimation of infants' expectations; (ii) justifying the selection of experimental parameters on clear theoretical grounds; and (iii) reporting pilot data when these are used to select experimental parameters, in order to make it clear that the selection is not based on a best-results criterion.

6. Eye-tracking solutions

In line with previous reviews (e.g., Aslin, 2007; Gredebäck, Johnson, & von Hofsten, 2009; Jackson & Sirois, 2009), I want to end this methodological review with a plea for the use of multiple eye-tracking measures in infancy research. In most VoE studies to date, infants' looking behaviour was recorded by a human observer who was looking at the infant through peepholes in the apparatus or from an adjacent room via closed-circuit video, and who pressed a button when the infant looked at the scene and away. A reliable cue to detect that an infant is looking away from the scene is to follow their head movement. However, it is in principle possible that an infant looks away without necessarily moving their head, in which case the issue arises as to whether a hidden observer would always be able to detect such fixations (e.g., did the baby look away from the scene or just towards the edge of the scene?).

Corneal reflection eye-tracking (as performed by modern eye-tracking systems) has a higher spatial and temporal resolution than the infant-looks-ahead/infant-looks-away observations that may be recorded by a concealed experimenter. More precisely, corneal reflection eye-tracking allows recording where in the scene an infant is fixating at any one time, and when and for how long an infant is looking away from the scene. Since these are important eye-tracking measures when trying to determine what an infant is paying attention to in a scene and when they may have lost interest in an event, the use of corneal reflection eye-tracking would significantly increase the accuracy of the measures of surprise used in VoE studies. Moreover, showing infants 'live performances' of events is not incompatible with using eye-tracking equipment: recent studies have used a new set-up that gives infants the impression that the eye-tracker monitor is a window with a real person behind (Hepach, Vaish, & Tomasello, 2015; Hepach, Vaish, & Tomasello, 2012).

Regarding false-belief research using the VoE paradigm, the use of corneal reflection eye-tracking would allow combining three different looking measures in a standard change-of-location task: first, measuring infants' anticipatory looking would allow establishing where they anticipate a mistaken protagonist will look for the target object, as shown by previous false-belief studies (He, Bolz, & Baillargeon, 2012; Southgate, Senju, & Csibra, 2007). Once the protagonist looks for the object in a certain location, infants' overall looking time could be used to determine a possible violation of their expectations (depending on whether the protagonist searched in one or the other container). Thirdly, the measurement of pupil dilation offers a promising way of detecting violations of expectation within participants (rather than between groups) and with a much higher temporal resolution (see, e.g., Gredebäck &

Melinder, 2010; Hepach et al., 2012; Jackson & Sirois, 2009; Laeng, Sirois, & Gredebäck, 2012; Sirois & Jackson, 2012). Regarding the importance of combining various eye-tracking measures, it is worth noting that Jackson and Sirois (2009) report task-evoked pupillary dilation (i.e. their measure of surprise was time-locked to when things ought to be surprising in the events), but this measure did not correlate with the typical looking-time measure they contrasted it with.

While eye tracking with infants poses its own technical challenges (e.g., eye detection and calibration), there is no reason why multiple eye-tracking measures could not be combined in the same study. In this sense, research on infant cognition need not continue to exclusively rely on gaze-tracking techniques that were devised more than 30 years ago, instead of taking advantage of the latest eye-tracking technology. In fact, given the necessarily indirect measures that must be used to investigate the cognitive development of pre-verbal infants (as opposed to that of children and adolescents, for example), it is all the more pressing that the most comprehensive and accurate measures of eye-tracking be combined in this challenging research field.

Acknowledgements

This research was supported by a *Young Research Talent Grant* from the Research Council of Norway (Ref. 230718). The author gratefully acknowledges this funding. Thanks to Chris Cummins, Bart Geurts, Julian Jara-Ettinger, Richard Moore and Hannes Rakoczy for their comments on earlier versions of this paper.

References

Aslin, R. N. (2007). What's in a look? *Developmental Science, 10*(1), 48–53.
Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. *Cognitive Development* (in press).
Barrett, H. C., Broesch, T., Scott, R. M., He, Z., Baillargeon, R., Wu, D., ... Laurence, S. (2013). Early false-belief understanding in traditional non-Western societies. *Proceedings of the Royal Society of London B: Biological Sciences, 280*(1755), 20122654.
Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language, 28*(5), 606–637.
Csibra, G., Bıró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science, 27*(1), 111–133.
Falck, A., Brinck, I., & Lindgren, M. (2014). Interest contagion in violation-of-expectation-based false-belief tasks. *Frontiers in Psychology, 5*.
Fenici, M. (2014). A simple explanation of apparent early mindreading: Infants' sensitivity to goals and gaze direction. *Phenomenology and the Cognitive Sciences, 14*(3), 497–515.
Gelman, A., & Loken, E. (2013). *Fishing expedition.* Department of Statistics, Columbia University.
Gelman, A., & Loken, E. (2014). The statistical crisis in science: Data-dependent analysis – A "garden of forking paths" – Explains why many statistically significant comparisons don't hold up. *American Scientist, 102*(6), 460–465.
Gergely, G., Nádasdy, Z., Csibra, G., & Biró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*(2), 165–193.
Gredebäck, G., & Melinder, A. (2010). Infants' understanding of everyday social interactions: A dual process account. *Cognition, 114*(2), 197–206.
Gredebäck, G., Johnson, S., & von Hofsten, C. (2009). Eye tracking in infancy research. *Developmental Neuropsychology, 35*(1), 1–19.
Grosse Wiesmann, C., Friederici, A. D., Singer, T., & Steinbeis, N. (2017). Implicit and explicit false belief development in preschool children. *Developmental Science, 20*(5), e12445.
He, Z., Bolz, M., & Baillargeon, R. (2011). False-belief understanding in 2.5-year-olds: Evidence from violation-of-expectation change-of-location and unexpected-contents tasks. *Developmental Science, 14*(2), 292–305.
He, Z., Bolz, M., & Baillargeon, R. (2012). 2.5-year-olds succeed at a verbal anticipatory-looking false-belief task. *British Journal of Developmental Psychology, 30*(1), 14–29.
Hepach, R., Vaish, A., & Tomasello, M. (2012). Young children are intrinsically motivated to see others helped. *Psychological Science, 23*(9), 967–972.
Hepach, R., Vaish, A., & Tomasello, M. (2015). Novel paradigms to measure variability of behavior in early childhood: Posture, gaze, and pupil dilation. *Frontiers in Psychology, 6*.
Hespos, S. J., Saylor, M. M., & Grossman, S. R. (2009). Infants' ability to parse continuous actions. *Developmental Psychology, 45*(2), 575.
Heyes, C. (2014). False belief in infancy: A fresh look. *Developmental Science, 17*(5), 647–659.
Jackson, I., & Sirois, S. (2009). Infant cognition: Going full factorial with pupil dilation. *Developmental Science, 12*(4), 670–679.
Kammermeier, M., & Paulus, M. (2017). Do action-based tasks evidence false-belief understanding in young children? *Cognitive Development* (in press).
Király, I., Jovanovic, B., Prinz, W., Aschersleben, G., & Gergely, G. (2003). The early origins of goal attribution in infancy. *Consciousness and Cognition, 12*(4), 752–769.
Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science, 330*(6012), 1830–1834.
Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2017). How robust are anticipatory looking measures of Theory of Mind? Replication attempts across the lifespan. *Cognitive Development* (in press).
Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is implicit Theory of Mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science, 29*(6), 888–900.
Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry a window to the preconscious? *Perspectives on Psychological Science, 7*(1), 18–27.
Luo, Y. (2011). Do 10-month-old infants understand others' false beliefs? *Cognition, 121*(3), 289–298.
Luo, Y., & Baillargeon, R. (2005). Can a self-propelled box have a goal? Psychological reasoning in 5-month-old infants. *Psychological Science, 16*(8), 601–608.
Luo, Y., & Baillargeon, R. (2007). Do 12.5-month-old infants consider what objects others can see when interpreting their actions? *Cognition, 105*(3), 489–512.
Luo, Y., & Johnson, S. C. (2009). Recognizing the role of perception in action at 6 months. *Developmental Science, 12*(1), 142–149.
Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255–258.
Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science, 308*(5719), 214–216.
Peterson, D. (2016). The baby factory: Difficult research objects, disciplinary standards, and the production of statistical significance. *Socius, 2*.
Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2017). Replications of implicit Theory of Mind tasks with varying representational demands. *Cognitive Development* (in press).
Rakoczy, H. (2015). In defense of a developmental dogma: Children acquire propositional attitude folk psychology around age 4. *Synthese,* 1–19 First Online.
Rubio-Fernández, P. (2018). What do failed (and successful) replications with the Duplo task show? *Cognitive Development* (in press).
Ruffman, T. (2014). To belief or not belief: Children's theory of mind. *Developmental Review, 34*(3), 265–293.
Scott, R. M. (2017). Surprise! 20-month-old infants understand the emotional consequences of false beliefs. *Cognition, 159*, 33–47.
Scott, R. M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs about object identity at 18 months. *Child Development, 80*(4), 1172–1196.
Scott, R. M., & Baillargeon, R. (2013). Do infants really expect agents to act efficiently? A critical test of the rationality principle. *Psychological Science, 24*(4), 466–474 PubMed Central ID: PMC3628959.
Scott, R. M., Baillargeon, R., Song, H. J., & Leslie, A. M. (2010). Attributing false beliefs about non-obvious properties at 18 months. *Cognitive Psychology, 61*(4), 366–395.

Scott, R. M., He, Z., Baillargeon, R., & Cummins, D. (2012). False-belief understanding in 2.5-year-olds: Evidence from two novel verbal spontaneous-response tasks. *Developmental Science, 15*(2), 181–193.
Scott, R. M., Richman, J. C., & Baillargeon, R. (2015). Infants understand deceptive intentions to implant false beliefs about identity: New evidence for early mentalistic reasoning. *Cognitive Psychology, 82*, 32–56.
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.
Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better *P*-curves: Making *P*-curve analysis more robust to errors, fraud, and ambitious *P*-hacking, a Reply to Ulrich and Miller. *Journal of Experimental Psychology: General, 144*(6), 1146–1152.
Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534.
Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*(6), 666–681.
Sirois, S., & Jackson, I. (2007). Social cognition in infancy: A critical review of research on higher order abilities. *European Journal of Developmental Psychology, 4*(1), 46–64.
Sirois, S., & Jackson, I. R. (2012). Pupil dilation and object permanence in infants. *Infancy, 17*(1), 61–78.
Song, H. J., & Baillargeon, R. (2008). Infants' reasoning about others' false perceptions. *Developmental Psychology, 44*(6), 1789–1795.
Song, H. J., Onishi, K. H., Baillargeon, R., & Fisher, C. (2008). Can an agent's false belief be corrected by an appropriate communication? Psychological reasoning in 18-month-old infants. *Cognition, 109*(3), 295–315.
Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science, 18*(7), 587–592.
Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science, 18*(6), 580–586.
Träuble, B., Marinović, V., & Pauen, S. (2010). Early Theory of Mind competencies: Do infants understand others' beliefs? *Infancy, 15*(4), 434–444.
Wagenmakers, E., Wetzels, R., Borsboom, D., van der Maas, H., & Kievit, R. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632–638.
Wang, S. H., & Baillargeon, R. (2005). Inducing infants to detect a physical violation in a single trial. *Psychological Science, 16*(7), 542–549.
Wellman, H. M. (2014). *Making minds: How theory of mind develops.* Oxford: Oxford University Press.
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.
Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*(1), 1–34.
Wynn, K. (1992). Addition and subtraction by human infants. *Nature, 358*(6389), 749–750.
Yong, E. (2012). Replication studies: Bad copy. *Nature, 485*(7398), 298–300.
Yott, J., & Poulin-Dubois, D. (2012). Breaking the rules: Do infants have a true understanding of false belief? *British Journal of Developmental Psychology, 30*(1), 156–171.
Yott, J., & Poulin-Dubois, D. (2016). Are infants' Theory of Mind abilities well integrated? Implicit understanding of intentions, desires, and beliefs. *Journal of Cognition and Development, 17*(5), 683–698.