Cognitive Psychology 91 (2016) 124–149

Contents lists available at ScienceDirect

Cognitive Psychology

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

Bayesian change-point analysis reveals developmental change in a classic theory of mind task

Sara T. Baker [a,b,c,*], Alan M. Leslie [a], C.R. Gallistel [a], Bruce M. Hood [b]

[a] *Department of Psychology and Center for Cognitive Science, Rutgers University, 152 Frelinghuysen Road, Piscataway, NJ 08854, USA*
[b] *School of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1TU, UK*
[c] *Faculty of Education, University of Cambridge, 184 Hills Road, Cambridge CB2 8PQ, UK*

ARTICLE INFO

Article history:
Accepted 18 August 2016
Available online 20 October 2016

Keywords:
Change point analysis
Bayesian statistics
Theory of mind
Preschoolers

ABSTRACT

Although learning and development reflect changes situated in an individual brain, most discussions of behavioral change are based on the evidence of group averages. Our reliance on group-averaged data creates a dilemma. On the one hand, we need to use traditional inferential statistics. On the other hand, group averages are highly ambiguous when we need to understand change in the individual; the average pattern of change may characterize all, some, or none of the individuals in the group. Here we present a new method for statistically characterizing developmental change in each individual child we study. Using false-belief tasks, fifty-two children in two cohorts were repeatedly tested for varying lengths of time between 3 and 5 years of age. Using a novel Bayesian change point analysis, we determined both the presence and—just as importantly—the absence of change in individual longitudinal cumulative records. Whenever the analysis supports a change conclusion, it identifies in that child's record the most likely point at which change occurred. Results show striking variability in patterns of change and stability across individual children. We then group the individuals by their various patterns of change or no change. The resulting patterns provide scarce support for sudden changes in competence and shed new light on the concepts of "passing" and "failing" in developmental studies.

[*] Corresponding author at: Faculty of Education, University of Cambridge, 184 Hills Road, Cambridge CB2 8PQ, UK.
E-mail address: stb32@cam.ac.uk (S.T. Baker).

http://dx.doi.org/10.1016/j.cogpsych.2016.08.001
0010-0285/© 2016 The Authors. Published by Elsevier Inc.

1. Introduction

Change is a ubiquitous yet recalcitrant problem for psychology. We would like to understand processes of change in many different areas from changes due to therapeutic or pedagogical interventions to natural changes wrought by learning and development. Yet all too often we are limited to comparing 'before and after snapshots' and filling the gap with speculation, or worse, with unexamined assumption.

Developmental change in children is often studied by comparing performance in cross-sectional time slices (between subjects) or by repeated longitudinal testing of the same individuals over lengthy periods. In either case, the study of change has relied on group data and this methodological constraint has served to forge theories of groups rather than of individuals (Estes, 1956). Group data are used because of the necessity of drawing statistical conclusions based on sufficiently large samples. Unfortunately, however, there is no such thing as a group brain. There are only individual brains wherein developmental and learning changes occur.

As Estes (1956) pointed out long ago, we cannot derive valid generalizations about the course of change in individuals from the course of change in a group average. The context for Estes' remarks were the many studies of learning in rats and other animals that show the familiar smooth, gradually incrementing learning curves predicted by associative learning theory that reflect the gradual strengthening of responses from baseline to asymptote as a function of number of learning trials. The problem with this picture, Estes pointed out, is that the gradual learning curve required by associative theory is seen only in the group curve averaged across many individuals. The curves of individual animals, by contrast, showed that learning was not gradual but occurred with sudden change. More recently, Gallistel, Fairhurst, & Balsam (2006) confirmed and quantified the abruptness of onset of conditioned learning in individual animals.

The first obstacle to understanding processes of developmental and many other kinds of change is that we typically lack even basic facts about what change looks like in the locus of change, the individual. We begin by confronting the fact, long recognized, but seldom addressed, that group-averaged data tells us almost nothing about variable developmental profiles of the individuals who make up the studied group. We then describe a method whereby repeated measures collected longitudinally from the same individual are combined with a new method of analysis, which we call, *Bayesian change-point analysis*. Each child in a cohort is repeatedly tested on a task in order to derive a cumulative record for that child of their performance over the period of testing. Cumulative records are ideal because each point in the record is a summary of past performance up to and including that point, while any change in performance produces a change in the slope of the curve. We then use a recently developed statistical method to discover and identify any point of change there may be in the record, which we call, *Bayesian change-point analysis* (Gallistel et al., 2004; Papachristos & Gallistel, 2006). The method of analysis is Bayesian; this has the important advantage of allowing us to demonstrate statistically not only points of change but also patterns of *no change* (Gallistel, 2009). This analysis yields an inferential statistic, the Bayes Factor, for a given individual's record without the need for group averaging. The Bayes Factor reflects the relative fit of a model to the individual's profile. Using the Bayes Factor as our guide, we are able to test the relative fit of two contrasting developmental models: either a change occurred in the individual's performance, or a change did not occur in the individual's performance. In other words, we can quantify our confidence that change occurred for an individual record. Just as importantly, for a given individual's record, the Bayes Factor will quantify our confidence that change did *not* occur, allowing us to distinguish *both* change *and* no-change from the case where an individual's record was simply uninformative. Additionally, this analysis will identify the point or points in the record where, with maximum likelihood, change or changes actually occurred.

Finally, to assess the replicability and generalizability of the methods and findings, we present two cohorts of single-case studies, from the US and the UK. By collecting a cohort of such single-case studies, we can derive unambiguous group descriptions based upon well-characterized individual cases without obscuring any individual differences. Here we apply these methods to the classic theory of mind shift in preschoolers. We reveal for the first time what performance looks like when this shift occurs, as well as when it does not occur and when, after lengthy periods of testing, performance records remain uninformative on this question. We then draw some tentative conclusions about the underlying processes of change. Our larger hope is that the methods presented here can supplement and add to existing methods of longitudinal data analysis in development.

We illustrate the single-case method in the domain of theory of mind development, using the much-studied "Sally and Anne" false belief task (Baron-Cohen, Leslie, & Frith, 1985). Representing people's thoughts and feelings, even when they are different from one's own, and using these representations of people's mental states to explain and predict people's actions, constitutes what is referred to as *theory of mind* (Premack & Woodruff, 1978). A case in point is the much-studied developmental change that takes place around the fourth birthday in untutored typically developing preschoolers in their ability to report correctly on the different perspective of another person who has a false belief. Despite nearly 40 years of intensive study, the causes and the processes underlying this change remain controversial and largely obscure (Gopnik & Wellman, 2012; Mahy, Moses, & Pfeifer, 2014). Furthermore, change in this area of cognitive development is not reserved for the period around the fourth birthday. A recent wave of findings shows that children much younger than three years—toddlers and even preverbal infants—can succeed when the tasks and behavioral measures are non-verbal (Wang & Leslie, 2016; Onishi & Baillargeon, 2005); moreover, development continues into adolescence (Devine & Hughes, 2013; Friedman & Leslie, 2004a, 2004b).

It may seem surprising to choose to study variability in a domain where change appears to occur on such a reliable time-scale. In theory of mind research, many group studies have shown that a mere 20–30% of typically developing three-year-olds are able to report a person's false belief when the person does not witness everything the child witnesses (Wimmer & Perner, 1983; see Wellman, Cross, & Watson, 2001, for a meta-analysis). We say that three-year-olds typically fail to attribute a false belief to the other person because *most* three-year-olds fail such a test. It is said that by the age of five, most children typically pass the false belief attribution task because group data show a high rate of perspective taking among five-year-olds, most averages approximating 80%. This cross-sectional result demonstrating a shift around the age of four has been replicated in many different cultural and socio-economic environments (e.g., Hughes & Cutting, 1999; Pears & Moses, 2003; Sabbagh, Xu, Carlson, Moses, & Lee, 2006). To put forward an account of theory of mind development implies that we can explain how children move from consistently failing to consistently passing such tests.

It is an empirical question how representative the group averages are of individual children (e.g., Carpenter, Call, & Tomasello, 2002; Nesselroad, Gerstorf, Hardy, & Ram, 2007). Without stopping to think about it one might assume that group profiles are representative of the underlying profiles of the individuals who make up the group. Actually, group rates of passing give us little information about individual rates of passing. For example, suppose we regularly find that 80% of five-year-olds pass a given task. Does that mean 80% of five-year-olds always pass, all will pass 80% of the time, half pass all the time and half pass 60% of the time, or 60% pass all the time and 40% pass 50% of the time, or any one of indefinitely many other patterns equally consistent with the group averages? We simply cannot tell from group data. If there is variability within an individual's record, is the observed variability due to low sensitivity in the instrument of measure (i.e., low reliability) or is it due to inherent cognitive vacillation? Indeed, the two may be confounded unless we are careful with our interpretations (Willett, 1989).

For developmentalists, the difference between an individual passing and failing is of paramount importance. Indeed, one aim of developmental research is to describe and explain how change occurs. What is the form of the transition between the failing and the passing states? Is it "sudden insight" or "gradual change" or something else? Do all children present the same profile of change from failing to passing? How can we characterize variability within and between individuals? These are the types of developmental questions that Bayesian change point analysis can answer.

1.1. Variability in theory of mind development

1.1.1. Inter-individual variability

Research in cognitive development tends to favor the dominant profile of behavior at the expense of a better understanding of inter-individual variability. We should question the implicit axiom that a picture of the group stands for a portrait of the individual. For example, one cannot conclude that smoothly decreasing reaction times in *group* data indicate that any given individual's reaction times are smoothly decreasing (e.g. Atance, Bernstein, & Meltzoff, 2010; Gallistel, Fairhurst, & Balsam, 2006). In an effort to characterize the developmental trajectory of theory of mind, Wellman and Liu (2004) applied a scalar model to group data to establish a suite of concepts, which children successively learn to explicitly manipulate. They gave children explicit mental and emotional state attribution tasks and found that they first became adept at attributing desires then types of beliefs followed by distinguishing real from apparent emotions. Still, the authors noted that approximately one in five children presented a developmental progression that was not consistent with the majority sequence. Most of these children were in the younger range of their sample. Similarly, Flynn (2006) derived a ranking of theory of mind tasks by difficulty based on group scores. When she examined individual children's performance on each task, however, less than one third of the individual children fit the sequence suggested by looking at group averages. That such a substantial number of individual records do not follow the trajectory implied by group data underscores the need to consider inter-individual variability as a source of information about conceptual development. Examining variability *between* individuals can inform our understanding of the cognitive mechanisms we wish to explain.

1.1.2. Intra-individual variability

The quest for generalization about the nature of cognitive change has also traditionally diverted attention from the complexity of conceptual development *within* individuals. Variability within individuals can be captured by testing them on more than one occasion with the same measure or at the same time with different measures. Regarding the classic theory of mind false belief task, the few studies that test individual children multiple times suggest that performance is often variable (Amsterlaw & Wellman, 2006; Flynn, 2006; Flynn, O'Malley, & Wood, 2004; Hughes et al., 2000; Mayes, Klin, Tercyak, Cicchetti, & Cohen, 1996). For example, Flynn (2006) examined children's performance on classic theory of mind tasks over a period of six months. She found very few cases where individuals simply went from failing 100% of the time to passing 100% of the time, with half of the improvements amounting to just one point on a scale of zero to seven. This reinforces the idea that intra-individual variability is pervasive and worthy of attention.

Furthermore, several lines of evidence show that different types of knowledge co-exist within an individual and develop separately (Church & Goldin-Meadow, 1986; Howe, Taylor Tavares, & Devine, 2014; Piaget, 1954; Zelazo, Frye, & Rapus, 1996). That is, variability in conceptual development within an individual can be recorded from one time point to the next, but also concurrently on different measures (e.g., tasks requiring explicit, implicit, verbal, visual, motor responses). Next we review evidence that calls into question the notion of general, unitary conceptual development at the level of the individual. Then, having established that inter- and intra-individual variability are more of a rule than an exception, we propose a new method to capture and quantify the variability that has hitherto been difficult to study.

1.1.3. Individual conceptual development

Research on infants' theory of mind has established an early tacit sensitivity to the mental states of others through looking time measures (Onishi & Baillargeon, 2005; for a review see Baillargeon, Scott, & He, 2010). During the preschool years children *fail* the explicit false belief attribution task, whereas measures of implicit knowledge and confidence levels point to a degree of understanding of the concepts at play (Ruffman, Garnham, Import, & Connolly, 2001). Thus, although group data from explicit measures may tell us that children below the age of four typically *fail* an explicit mental state attribution task, performance by the same individuals can be very different on implicit measures. Likewise, although group data indicate older children and adults typically *pass* explicit belief attribution tasks under most circumstances, there is more to the story than a sudden insight after which the conceptual content is mastered once and for all. Of course, cognitive development in theory of mind does not stop in early childhood. Implicit and explicit attribution of mental states continue to be measured distinctly one from another even in adulthood (e.g., Cohen & German, 2009; Keysar, Barr, Balin, & Brauner, 2000; Wang & Leslie, 2016).

Although longitudinal work suggests some degree of continuity between infants' tacit understanding and preschoolers' explicit attribution of mental states (Wellman, Lopez-Duran, LaBounty, & Hamilton, 2008), dissociations between response modalities are well documented in theory of mind and other domains of reasoning like physics and math (Clements & Perner, 1994; Goldin-Meadow, Alibali, & Church, 1993; Hood, Cole-Davies, & Dias, 2003; Howe, Tavares, & Devine, 2012). A child may fail on one measure, while passing on another. Hence when talking about conceptual development, it is not meaningful to speak in broad terms about ''failing'' or ''passing'' theory of mind tasks and the challenge we face is more than simply describing how one moves from one state to another. In reality states of knowledge are much more complex. Appropriate methods of data collection and analysis can help us to render a truly mechanistic developmental picture.

Thus far we have drawn on the literature to illustrate how conceptual development in theory of mind is not likely to be simple and monotonic. Even if development were monotonic for any given capacity under study, cross-sectional group averages are statistical artifacts that do not reflect individual growth trajectories. Indeed, cross-sectional group data give a schematic idea of the relative rates of passing in groups, but they cannot tell us about the individual's pathway through development. How can we go beyond the methodological limitations that constrain the theoretical frameworks we adopt?

1.2. How to study variability in developmental change

Intra-individual variability has been particularly difficult to study. How do we tell when an individual child has undergone developmental change? Equally important, how do we determine if a child has *not* undergone change? Showing no change is equivalent to proving the null hypothesis, something which standard significance testing cannot do. The method we introduce here solves these problems, allowing us to determine for each individual child when the evidence says a change in performance did or did not occur. Our method requires a rich set of longitudinal data from individuals, which can then be subjected to non-traditional statistical analyses. In the next sections we outline the intensive behavior sampling approach, known as the microgenetic method, which relies on a longitudinal research design. Then we summarize existing statistical techniques for analyzing change. Finally, we introduce the novel Bayesian change point algorithm adapted here from the animal learning literature.

1.2.1. Intensive data sampling throughout a period of change

Experimental work in child development has largely relied on group data to examine developmental change. In neuropsychology, in contrast, single case studies are standard practice. Despite being controversial, it is nevertheless recognized that the observation and analysis of individuals can inform the plausibility of general models (Caramazza, 1986; Glymour, 1994; Shallice & Evans, 1978). In psychophysical studies, too, (multiple) single case studies with repeated measures are the norm. In developmental psychology, Siegler has been an influential proponent of collecting data on individual children's performance throughout a developmental period (Siegler, 2007). He has consistently reminded us that change is not monotonic, nor uni-dimensional. He and his colleagues have been able to demonstrate this with experiments that are designed to accommodate individual variability. Crucially, rather than each child contributing only one data point to a study, Siegler's microgenetic method samples a child's behavior repeatedly throughout a transitional period, often uncovering wave-like patterns of task solutions. In various tasks, such as addition and categorization, children's capacities seem to ebb and flow before they settle (Siegler, 1987; Siegler & Svetina, 2002). Similar intra-individual variability has been found in motor development, indicating that the non-monotonic function is a more general characteristic of development, which extends beyond cognitive development (Adolph, Robinson, Young, & Gill-Alvarez, 2008). However, despite the richness of a microgenetic dataset, these studies use traditional inferential statistics based on group averages, such as ANOVA, or those noted below. Traditional group-based statistics do not exploit the strengths of intensive data sampling at the individual level (see Cheshire, Muldoon, Francis, Lewis, & Ball, 2007 within a special issue on the microgenetic method).

1.2.2. Analyzing individual profiles of development

In the case of theory of mind development, the few microgenetic studies that have been conducted have all been limited by available statistical analyses (Amsterlaw & Wellman, 2006; Flynn, 2006; Flynn et al., 2004). Even when figures show individual profiles, group statistics are used to draw conclusions about individual change (e.g., Amsterlaw & Wellman, 2006, p. 158). The dissonance between individual profiles and group-level analyses is more apparent in microgenetic research than in other longitudinal or cross-sectional designs. For example, Flynn et al. (2004) selected 21 preschool children who initially failed most of their tasks then tested them once per month for six months. Although as an averaged group they showed steady improvement, there were several different individual profiles, including nine children (43%) who began to pass the tasks, only to fail again later. Is such variability just statistical noise and thus meaningless or is it real and meaningful?

Traditional statistical techniques tend to highlight invariant features across individuals, at the expense of examining variability. Developmental questions have been addressed using statistical techniques such as backwards trial graphing (Opfer & Siegler, 2004), survival analysis (Singer & Willett, 1991), fuzzy sets (Van Dijk & Van Geert, 2007), linear and non-linear growth curve modeling (Grimm, Ram, & Hamagami, 2011; Jordan, Hanich, & Kaplan, 2003; Willett, 1989), non-normal random effects modeling (Agresti, 2000 in Cheshire et al., 2007), stability coefficients (Hoffman, Jacobs, & Gerras, 1992) and change scores (Ferrer & McArdle, 2010). Some of these methods to some degree address individual variability (e.g. growth curve modeling), but most statistical methods used by developmental psychologists ultimately rely on group datasets. Even statistical analyses that capture more richness in the data than frequently used group analyses sometimes highlight information about invariance over information about variance. For example, Rasch analysis is employed to identify the best fitting invariant sequence of development (Boom, Wouters, & Keller, 2007; Lamb, Vallett, & Annetta, 2014; Young et al., 2011). Individual and measurement variance are part of the input for Rasch modeling, but individual profiles are not the objective of the output. Instead, the analysis delivers one ordered sequence of items that represents, on average, the collection of sequences displayed by a set of individuals. Therefore Rasch analysis does not reveal variability within individual children's records. In general, analytic techniques that look more closely at intra- or inter-individual variability have been under-used.

At least part of the reason for the reliance on group data is a concern for the power of induction based on traditional null hypothesis significance tests (tNHST). Where the methods are available, we join a growing number of researchers in supplementing tNHST with the use of Bayesian approaches (e.g., Dienes, 2011; Kruschke, 2010). To make our case, we next outline our reasons for adopting a Bayesian change-point analysis.

1.3. Quantifying individuals' developmental change: Bayesian change point analysis

The Bayesian change-point analysis employed here differs from traditional techniques in developmental research in two main ways. First, it focuses on change at the level of the individual. The approach has its roots in the animal learning literature, where, as in child development, learning is typically assumed to follow the smoothly rising curve of averaged group data. However, work with pigeons and mice has revealed that this curve can be an artifact of averaging over individuals who show abrupt changes on earlier or later trials (Gallistel et al., 2004; Papachristos & Gallistel, 2006). Because learning, like development, takes place in the individual brain, understanding of the acquisition process should be based on individual, not group, curves.

The second main difference between the Bayesian change point analysis and traditional techniques in developmental research is that, as the name suggests, it uses Bayesian statistics, rather than tNHST. In contrast to tNHST, Bayesian change point analysis provides a tool for evaluating the strength of the evidence in favor of various models of change in the probability of obtaining a value in a chronological record (Dienes, 2011; Lee & Wagenmakers, 2005; Salsburg, 2001). That is, we can assess the fit of a model of change, alongside the fit of a model of no change, in the record of performance. A major reason for computing Bayes Factors rather than *p* values is that Bayes Factors may be against or in favor of a null hypothesis. Unlike *p* values, they can tell us how strongly the data favor the conclusion that there was no change. In the demonstration we provide here, this translates as modeling changes in the probability of a child passing the theory of mind task. Of course it could apply to modeling changes in a patient's heart rate over the course of pharmaceutical treatment, or changes in population density in bird ecology, or any other developmental question where data are plotted against time.

For the analysis, we use an algorithm that runs as MatLab code (for more information see Appendix A) to automatically evaluate the individual records of performance. The algorithm generates output to reflect the mathematical evaluations of model fit, and the researcher then interprets this output, much like traditional software packages for social science research. Interpretations are based around the so-called Bayes Factor, which gives information about the relative fit of models for a given set of data. The Bayes Factor, unlike the traditional *p*-value, is not subjected to a strict threshold above which results should be disregarded. Instead, the Bayes Factor more closely reflects human decision making, as it is a ratio of likelihoods. For instance, one model may be favored 3:1 over another model, which means that, given only the evidence of the data, one model is three times more likely than the other. It is up to the researcher to decide what conclusions to draw based upon this likelihood ratio. Conventions in Bayesian statistics in psychology generally suggest that a 3:1 ratio implies substantial evidence in favor of one model, a 10:1 ratio implies strong evidence, and so on (see Wetzels et al., 2011, for discussion).

Beyond the Bayes Factor, visual inspection of the cumulative records of performance, along with examination of the parameter estimates the algorithm returns (e.g. the point in the record where performance changed), are also useful to gauge the qualitative nature of the models of the data. Thus, the analytic strategy involves first, asking whether there is evidence in each individual record for a model of change or no change in the probability of passing the task at hand. Then, we can turn to the question of what kind of model provides the best fit (change or no change), and, if there are any points where the probability of passing does change, when these occur in the record, and what the probability of passing before and after this point was.

Finally, a note on what our Bayesian approach is not. It is not an attempt to model children's cognitive processes *per se*. The distinction between a psychological model and a mathematical model is of capital importance and is discussed in greater detail by others (see for example Bowers & Davis, 2011; Kruschke, 2010). Recent Bayesian approaches to the study of child development have suggested that children's reasoning itself may embody Bayesian principles (e.g., Bonawitz, Van Schijndel, Friel, & Schulz, 2012; Goodman, Ullman, & Tenenbaum, 2011; Gopnik & Wellman, 2012; Griffiths, Vul, & Sanborn, 2012; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; see also special issue of *Cognition*, vol. 120, 2011). While we remain open minded on this point, our present Bayesian modeling is purely descriptive and agnostic as to the cognitive processing mechanisms that give rise to the data we observe.

1.4. Bayesian change point analysis of the 'preschool shift' in theory of mind

To assess the feasibility of applying the Bayesian change point analysis to human development, we repeatedly tested preschool children on standard belief attribution tasks, deriving a cumulative record of each child's performance over the course of the study.

2. Method

2.1. Participants

We aimed our recruitment at the transitional phase of theory of mind development based on previous cross-sectional and longitudinal data. Our final sample consisted of 52 children from the US and the UK (32 girls, 20 boys; mean age at start: 47.3 months, sd: 6.3, range: 34–62). Although in theory the trials could be separated by any length of time, in practice it is difficult to interpret a learning curve with widely varying inter-session intervals. Given that the developmental time-course of the processes under investigation is largely unknown, a decision was made to only include children whose trials were not spaced more than six weeks apart. Ten additional children missed more than two consecutive sessions and were thus excluded from our analyses. Although the minimum number of trials

Table 1
Mean (standard deviation) for two cohorts of potential improvers and for initial passers.

	N	Start age (in months)	Number of trials	Overall probability of passing
Cohort 1 US	23	47.2 (6.7)	21.1 (8.8)	32.1 (29.3)
Cohort 2 UK	20	45.5 (5)	20.3 (6.9)	58.5 (32.5)
Initial passers	9	52.7 (4.8)	19.9 (7)	91.2 (10.2)

for change point analysis is not well established, we also excluded five additional children who completed less than five testing sessions (see Table 1 for final sample demographics). Because this was the first use of this type of analysis with humans, the criterion for a minimum number of data points was unknown. For this reason, and unlike previous microgenetic investigations of theory of mind, we set out to test children for as long as possible, without limiting the duration of or the number of sessions in the study. Individuals in the final sample completed 21 trials on average (sd = 7.6). Testing sessions occurred approximately monthly (mean interval between sessions = 28 days, sd = 11). All participants had written consent from guardians prior to the beginning of the project and gave verbal assent prior to taking part in each session.

2.2. Materials

We used classic true and false belief attribution tasks (Baron-Cohen et al., 1985; Hogrefe, Wimmer, & Perner, 1986). A major concern in longitudinal research is measuring the same construct repeatedly, which can lead to practice effects interfering with the natural course of development. In this age group the false belief transfer task and the false belief contents task have been shown to be of equivalent difficulty (Wellman et al., 2001). We therefore used these interchangeably to increase the number of trials without jeopardizing the validity of the measurements. In addition, specific materials changed for each task and for each session so that children were not able to build a stimulus-response association for the tasks.

2.3. Procedure

Every testing session for both cohorts was conducted by the same female experimenter and took place either in a quiet room of the preschool, in a laboratory setting or in the child's home. Each session lasted about 15 min. Twenty-two children completed one trial of false belief unexpected transfer (Baron-Cohen et al., 1985) and one trial of false belief unexpected contents (Hogrefe et al., 1986) in each session. Twenty-five children completed two trials of false belief unexpected transfer in each session and one trial of false belief unexpected contents in alternate sessions. Five children completed two trials of false belief unexpected transfer and two trials of true belief expected transfer in each session.

In the transfer tasks a character (e.g., Sally) put an object in a hiding location before leaving the scene. Then, either (a) Sally returned to see the object being transferred to another location (expected transfer leads to character's true belief about the location of the object), or (b) Sally returned after the object had been transferred to another hiding location (unexpected transfer leads to character's false belief about the location of the object). The critical test question required the child to predict where Sally would look for the object upon her return, or where she would think the object is. For false belief scenarios children received one point if they responded Sally thought the object was still in its original location, or if they predicted she would look there upon her return. All other responses were scored as zero. Scores were reversed for the true belief scenario (see Appendix B for details of scenarios).

In the unexpected contents task, children were shown a container and then discovered that it contained an unexpected object. They were then asked to predict what another person would think is inside, having never seen the contents. They received one point if they said another person would think the container concealed the expected rather than actual object. All other responses were scored as zero.

Each task also included two control (reality and memory) questions. The order of tasks, test questions and control questions was counterbalanced across sessions. No feedback was given. At the end of each session, children received three stickers.

3. Results

3.1. Data treatment: Bayesian change point algorithm

The algorithm runs as Matlab code (Math Works, 2000; for more information see Appendix A). The algorithm proceeds for each record separately by a process of elimination, evaluating models in turn and producing as output the model of the data garnering the strongest evidence. In our illustrative example, we were atheoretical with respect to model selection (i.e., we did not specify a particular model to test), as no theory of theory of mind development makes a strong prediction about the shape of developmental change. Therefore, a preliminary analysis checked to see whether any of the data sequences of individual performance could be best represented by a model in which there was more than one change in the probability of a correct answer. In none of the records was this the case, meaning that all of our individual records could be modeled with one change or no change in the probability of passing.

3.2. Bayes Factors

We computed the Bayes Factor (BF) for the contrast between a model of no change in the record of performance and a model with one change in the record of performance, given even prior odds of a change in a record of that length. The Bayes Factor is the ratio of the marginal likelihoods. It is also the factor by which the data change the prior odds in favor of one model or another. When $BF = 1$, the odds are equal in favor of both the null and alternative hypothesis, in other words, the data are equally likely under a model of the data implying no change (null) or change (alternative) in performance. By convention, when $1 < BF < 3$, we say that the evidence favoring one model over another is weak; when $3 < BF < 10$, we say the evidence is substantial and when $10 < BF$, we say the evidence is strong (Wetzels et al., 2011; Jeffreys, 1961). Researchers are invited to make their own judgments about the strength of the evidence they require to draw conclusions (unlike classic *p*-values which have evolved with cut-off points for 'significant' findings). We see this as an advantage because people have an intuitive grasp of "the odds favoring one hypothesis over another," so there is less need for artificial stipulations. As Gallistel (2009:445) puts it,

> "Telling the research community what odds are required to decide in favor of a hypothesis is a bit like telling the betting community what odds are required to make a safe bet. The odds are what they are."

3.3. Weight of the evidence

Another way to convey the extent to which the data favor one model over another is by the common log of the Bayes Factor, also called the weight of the evidence. Equivalent weights of opposite sign represent equivalent changes in the odds in opposite directions. A 0.5 change in weight either way represents a 3-fold change in the odds. The hypotheses we contrasted for each individual, were (a) that the probability of their passing changed in the course of testing, versus (b) that the probability of their passing did not change. One can see in Fig. 1 the weight of the evidence supporting a change versus a no-change model for each individual record. If the bar in Fig. 1 extends above 0, the odds in favor of a change are better than even; if below, worse than even (that is, the odds are against a change). The solid lines at -0.5 and 0.5 indicate likelihood or odds of 3.16:1 that the evidence supports the hypothesis that there was no change or supports the hypothesis that there was a change, respectively. We use these odds as Bayesian decision criteria, very roughly, as equivalent to the 5% alpha level of tNHST (Gallistel, 2009). Thus, a bar extending below the -0.5 line indicates at least substantial evidence in

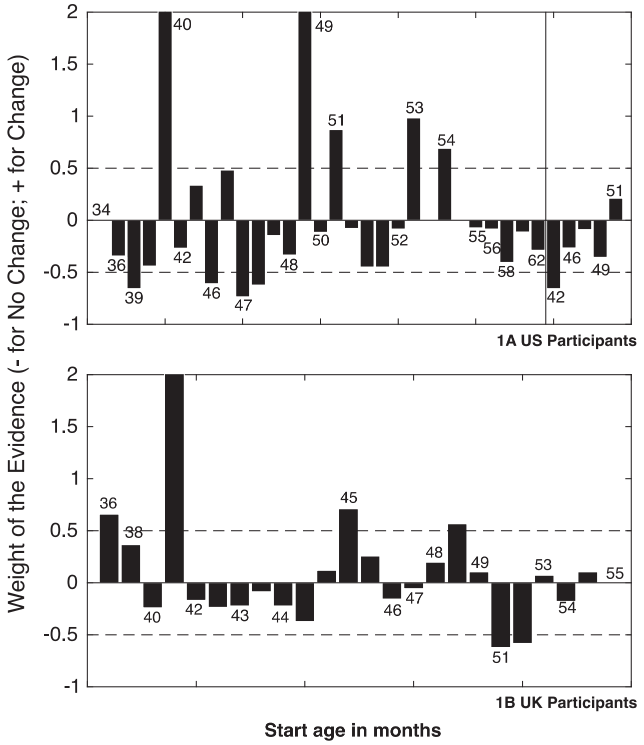

Fig. 1. Weight of the evidence in favor of change (positive) or no change (negative) for each participant's cumulative record. Records appear from left to right, from youngest (34 months, weight of evidence = 0) to oldest starting age (62 months). Dotted lines show weight of evidence equivalent to BF = 3.16. (A) US participants' false belief records; the rightmost five records reflect true belief tasks. (B) UK participants' false belief records.

favor of the no-change hypothesis, while a bar extending above the 0.5 line indicates at least substantial evidence for a change. In between these lines, the evidence in favor of one hypothesis over the other is only weakly informative.

It is important to bear in mind that results are deliberately situated at the individual rather than group level. The aim is to examine individual records and attempt to draw conclusions about the best model (change or no change) for each child studied, that is, for each of the 52 cumulative records of the number of correct responses in the false belief task and the five cumulative records of the number of correct responses in the true belief task. Fig. 1 shows each individual's Bayes Factor and associated weight of evidence in favor of change or no-change.

Having examined model fit (change or no change), our analytic strategy next focuses on the visual inspection of the individual cumulative records. Individual cumulative records provide a more detailed view of performance. In each record the x-axis represents successive trials and the y-axis represents the cumulative successes. Because children received and passed different numbers of trials in each session and overall, the axes differ between participants (see Section 2 for procedural details). Each time the child passed a trial, their cumulative record increased by one. Each time they failed, the cumulative record increased by zero. Thus, if the child performed consistently, the cumulative record would have a constant slope. Any change in the probability (p) of passing would imply a change of slope in the cumulative record. By convention, when the Bayes Factor is greater than 1 (i.e. the evidence favors one model over another), a change point is indicated by a vertical line at the estimated change point. The line is dashed if the Bayes Factor is less than 3 (weak evidence); it is solid when the Bayes Factor is greater than 3 (substantial evidence). When there is evidence of a change, the before and after estimates of p are given (p_1 and p_2). Otherwise the estimate of the stationary (unchanging) p is given. These stationary p's range from 0 to greater than .9, reflecting a wide range of performance across the cohort.

3.4. True belief control task

We first examined children's cumulative records of performance in the true belief task (Fig. 2). Visual inspection of the records reveals consistently high rates of passing in this control condition. However, the strength of the evidence allowing us to draw conclusions was weak in all cases save one (TB6), where there was substantial evidence in favor of a model of no change. The absence of evidence one way or the other when performance is near the upper limit is a ceiling effect: The no-change and the +change descriptions must be roughly equally likely in records of 30 trials or less, because it takes many trials for evidence of an improvement to accumulate once it has occurred.

3.5. False belief task

The following sections examine in further detail the records of false belief performance. The distribution of records reflected varying levels of overall performance throughout the course of testing (Fig. 3; see also Appendix C).

3.5.1. Initial passers

Nine children passed the first three trials (Table 1 and Fig. 4). These children were considered unlikely candidates for exhibiting the developmental change under investigation. Hence we present them as a separate category of 'initial passers' at the start of testing, in contrast with 'potential improvers' who failed at least one of the first three trials.

3.5.2. Potential improvers

The remaining 43 children who failed at least one of the first three trials ('potential improvers') were then considered (see Table 1 for descriptive statistics of the two cohorts of potential improvers).

We examined the records of performance for each type of model of the data according to the evidence resulting from the Bayesian change point analysis, and sorted the individual records into categories: substantial evidence of change or no-change (weight of the evidence >0.5, equivalent to

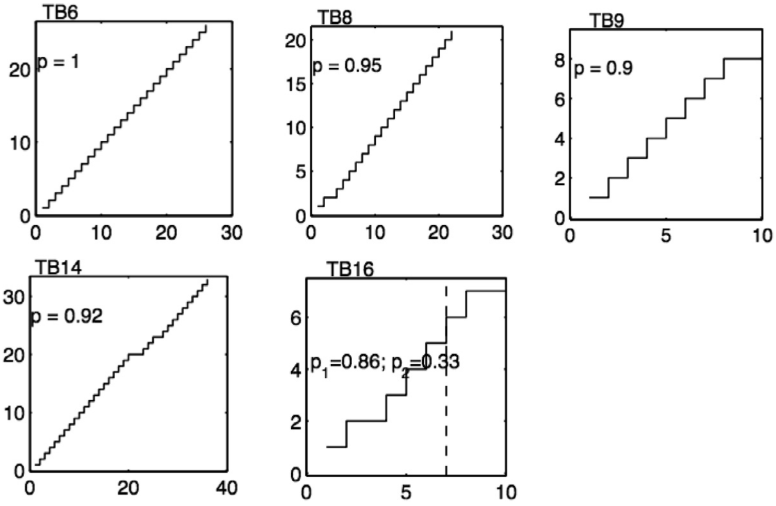

Fig. 2. Cumulative records of performance in the true belief task (n = 5 from US cohort). X-axis: number of trials; Y-axis: number of successes.

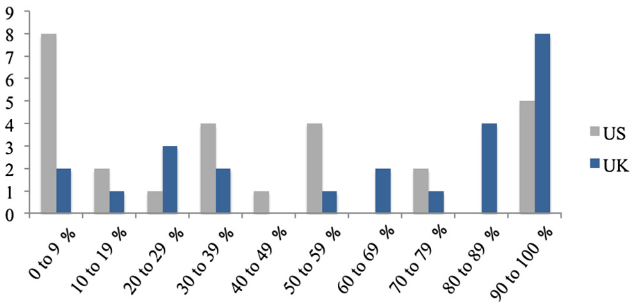

Fig. 3. Frequency of individual records of performance (y-axis) from the two cohorts as a function of the overall probability of passing the false belief task (x-axis) throughout the period of testing.

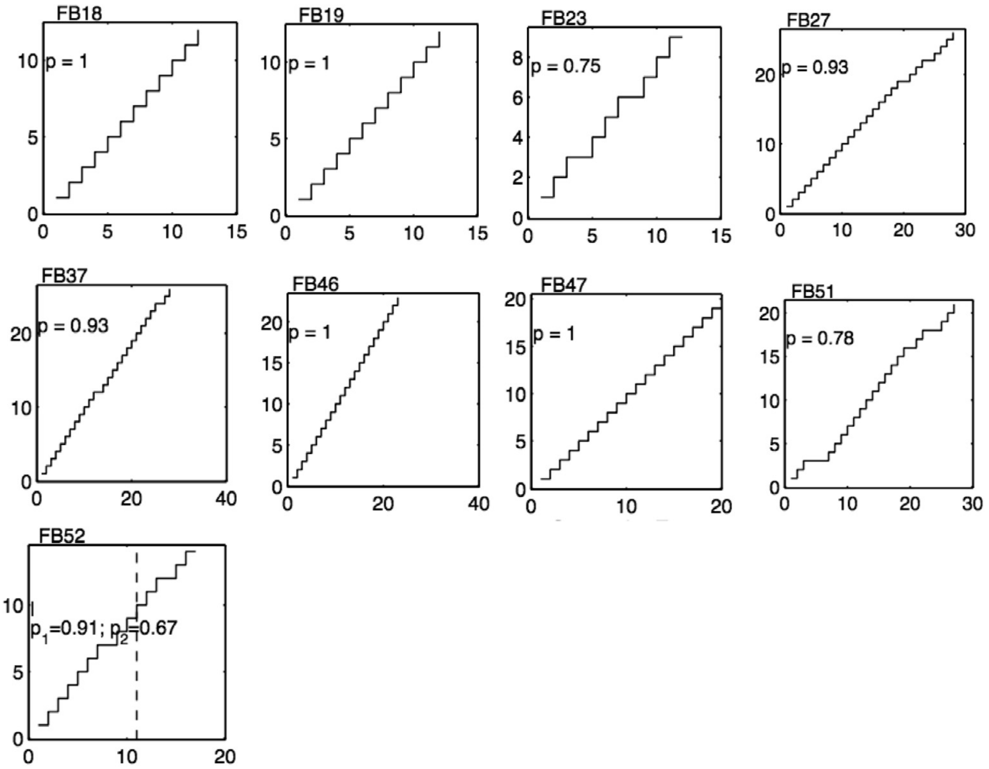

Fig. 4. Cumulative records of performance for initial-passer profiles at the start of testing (n = 4 US; n = 5 UK). X-axis: number of trials; Y-axis: number of successes. Note that initial-passer profiles were defined as those children who passed the first three consecutive trials.

3 < BF), weak evidence of change or no-change (weight of the evidence >0 and <.5, equivalent to BF < 3; nomenclature based on Jeffreys, 1961). A final category was constituted of two records where the weight of the evidence was close to 0 (BF close to 1, indicating data equally favor both models), barring any conclusions from being drawn.

Children from both the New Jersey and the Bristol cohorts were spread across these categories. Of 43 cumulative records of performance, about one third provided conclusive evidence for a model of change or no change in the probability of passing (see Tables 2a and 2b). The remaining two thirds provided only weak evidence one way or the other. Reasons for this are explored in Section 4.

Table 2a
Distribution of individual records of improvers from US cohort (n = 23), according to the evidence for change or no change.

	Substantial evidence for		Weak evidence for		Uninformative
	Change	No change	Change	No change	
Percentage of improvers (frequency)	21.7 (5)	17.4 (4)	8.7 (2)	47.8 (11)	4.3 (1)
Mean start age (in months)	49.4	44.8	44	48.8	34
Number of trials	22.4	29	16	18.6	22
Overall rate of passing (in %)	44.3	0.7	43.2	38	9.1
Bayes factor in favor of each model (range)	4.63–1811.5	3.57–4.76	2.12–2.83	1.19–2.56	1.01
Putative change point trial (mean)	12.4	–	7.5	–	–

Table 2b
Distribution of individual records of improvers from UK cohort, according to the evidence for change or no change.

	Substantial evidence for		Weak evidence for		Uninformative
	Change	No change	Change	No change	
Percentage of improvers (frequency)	20 (4)	0	35 (7)	40 (8)	5 (1)
Mean start age (in months)	42.3	–	47.3	43.4	54
Number of trials	25	–	17	20	26
Overall rate of passing (in %)	41.2	–	48.2	72.9	84.6
Bayes factor in favor of each model (range)	3.59–10,679	–	1.17–2.22	1.12–1.7	.99
Putative change point trial (mean)	14.8	–	8	–	–

Regarding the shape and rate of change, the diversity of profiles is striking, and has never been quantified before at the level of the individual. Among records with substantial evidence for change, three records present a pattern akin to 'sudden insight' (FB31, FB5, and FB 16, plus maybe FB14 and FB39; Fig. 5). The rest show other patterns, including change from 12% to 50% passing (FB44) and change from 40% to 100% (FB22). Similar diversity of change patterns can be observed in the records presenting only weak evidence for change.

Where there is substantial evidence for a no change model, we only observed records with floor performance (Fig. 6; recall that profiles presenting ceiling performance at the start of testing were deemed 'initial passers' and thus not included in subsequent considerations).

In many ways the 19 records presenting weak evidence for a model with no change in performance are the most interesting (Fig. 8). In this group we observe 10 records (53% of this category) with stably unstable performance between 10% and 90% overall passing rate. The fact that children can perform so inconsistently over several months could not be revealed with group-based analyses.

3.5.3. Strength of the evidence

It is methodologically informative to compare records with substantial versus weak evidence for change (Figs. 5 and 7) and for no change (Figs. 6 and 8). Whereas there are equal numbers of records in the change groups garnering substantial and weak evidence, there is a much larger number of records in the no change group garnering weak evidence, compared to substantial evidence. We return to some mathematical reasons for why this should be so in Section 4 (see Fig. 9).

The overall number of trials is greater for records garnering substantial evidence (whether for or against change). However, at the level of the group there was no appreciable correlation between number of trials and strength of the evidence (Kendall's *tau* (43) = −.093). One indicative difference between these records is that the putative change point for profiles with merely weak evidence for change occurs much earlier compared to the profiles with substantial evidence for change (see Table 2). This suggests that a lengthy baseline of consistent performance may be desirable in order to detect changes in the probability of passing a task.

3.5.4. Age as a factor in predicting developmental change

Age did not predict which category of model children's records fell into. Profiles with substantial evidence for change correspond to young three-year-olds all the way through to four-and-half-

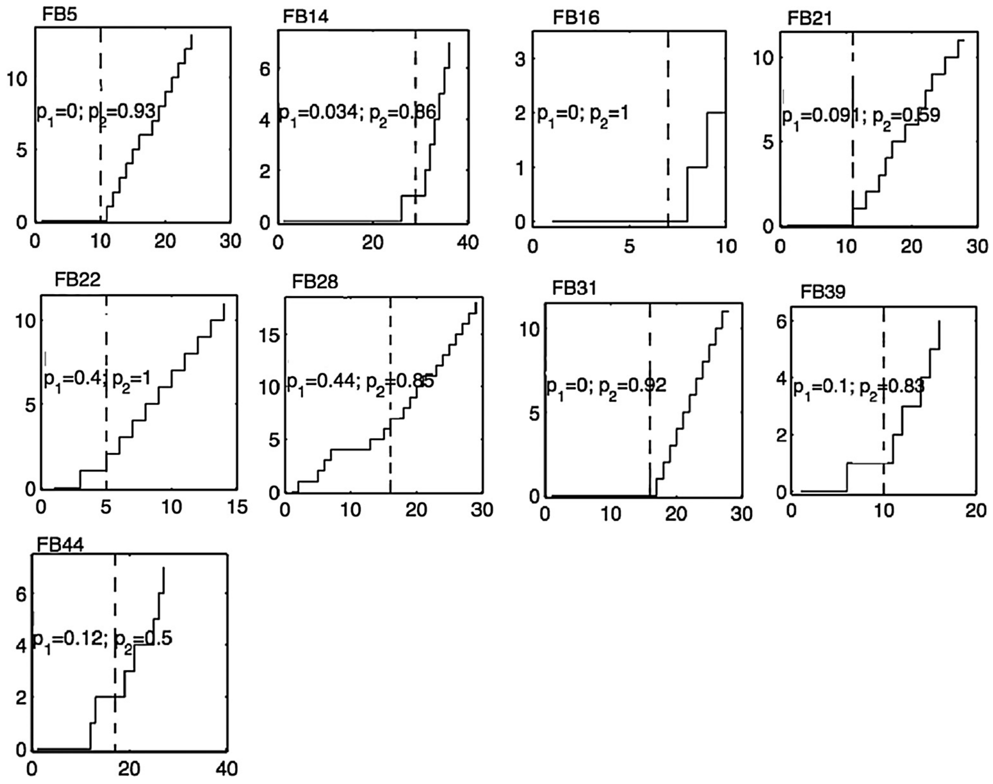

Fig. 5. Substantial evidence for a model of change (n = 5 US; n = 4 UK). X-axis: number of trials; Y-axis: number of successes.

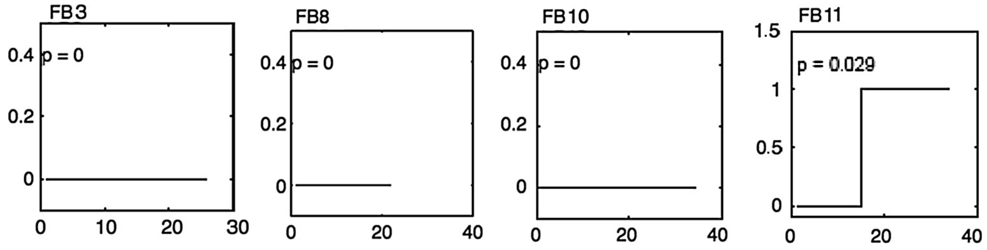

Fig. 6. Substantial evidence for a model of no change (n = 4 US). X-axis: number of trials; Y-axis: number of successes.

year-olds at the start of testing (Fig. 5). Although the mean age at the stipulated change point is 50 months, this varied widely from child to child (sd = 6 months). Records with substantial evidence for no change (Fig. 6) also correspond to children of varying ages (range = 39–47 months at start). These profiles all represent floor performance throughout several months of testing.

4. Discussion

4.1. Variability in developmental profiles

We began by pointing out the ambiguities inherent in traditional group averaged measures of the age of ‘success’ on the false belief task. We illustrated how the Bayesian change point analysis may be

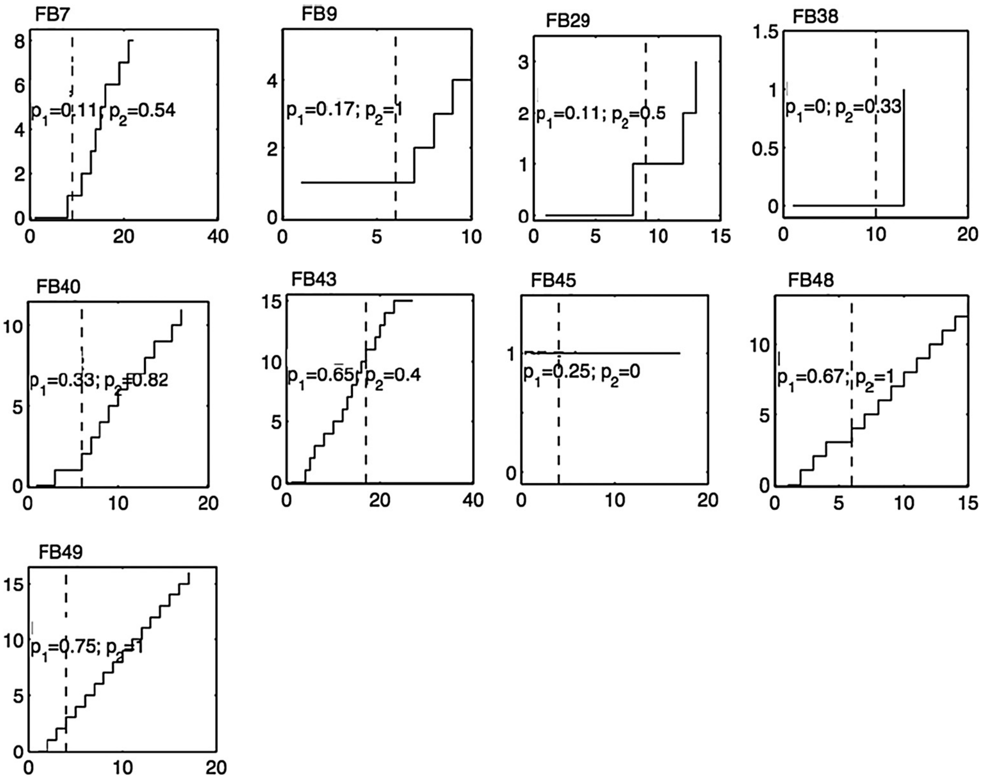

Fig. 7. Weak evidence for a model of change (n = 2 US; n = 7 UK). X-axis: number of trials; Y-axis: number of successes.

used to address some theoretical and methodological concerns with traditional studies. Our results from multiple single case studies in two countries underline these hazards. There is no single answer to the question, what do the traditional group averages mean? Some individuals maintain consistent passing and some maintain consistent failing over long periods; some individuals show 'sudden' change while many individuals pass through long periods of variable performance. If one really wishes to characterize developmental patterns, one must begin by accurately characterizing individual cases. Take, for example, participants FB5 and FB24 (Figs. 5 and 8). Their overall mean probabilities of passing the false belief task are almost identical, around 50% passing rate, traditionally regarded as 'chance' responding. Yet Bayesian analysis of the individual cumulative records reveals two radically different developmental profiles, one showing 'sudden' change from floor to ceiling, the other a kind of 'stable instability'. Our results and analyses provide some insight into the individual profiles behind group averages in the theory of mind literature. It is important to bear in mind that, unlike traditional significance testing, a Bayesian analysis allowed us to evaluate a hypothesis of no change as well as a hypothesis of change (Gallistel, 2009).

4.2. On the shape and rate of change

We found a striking variety of developmental patterns (c.f. Siegler, 1987). When considering records of performance in both false belief and true belief tasks, 14 records show substantial evidence either for change (nine children) or for no change (five children). Even when the evidence is weak, we are able to quantify the balance of evidence favoring change or no change in nearly every case (see Table 3).

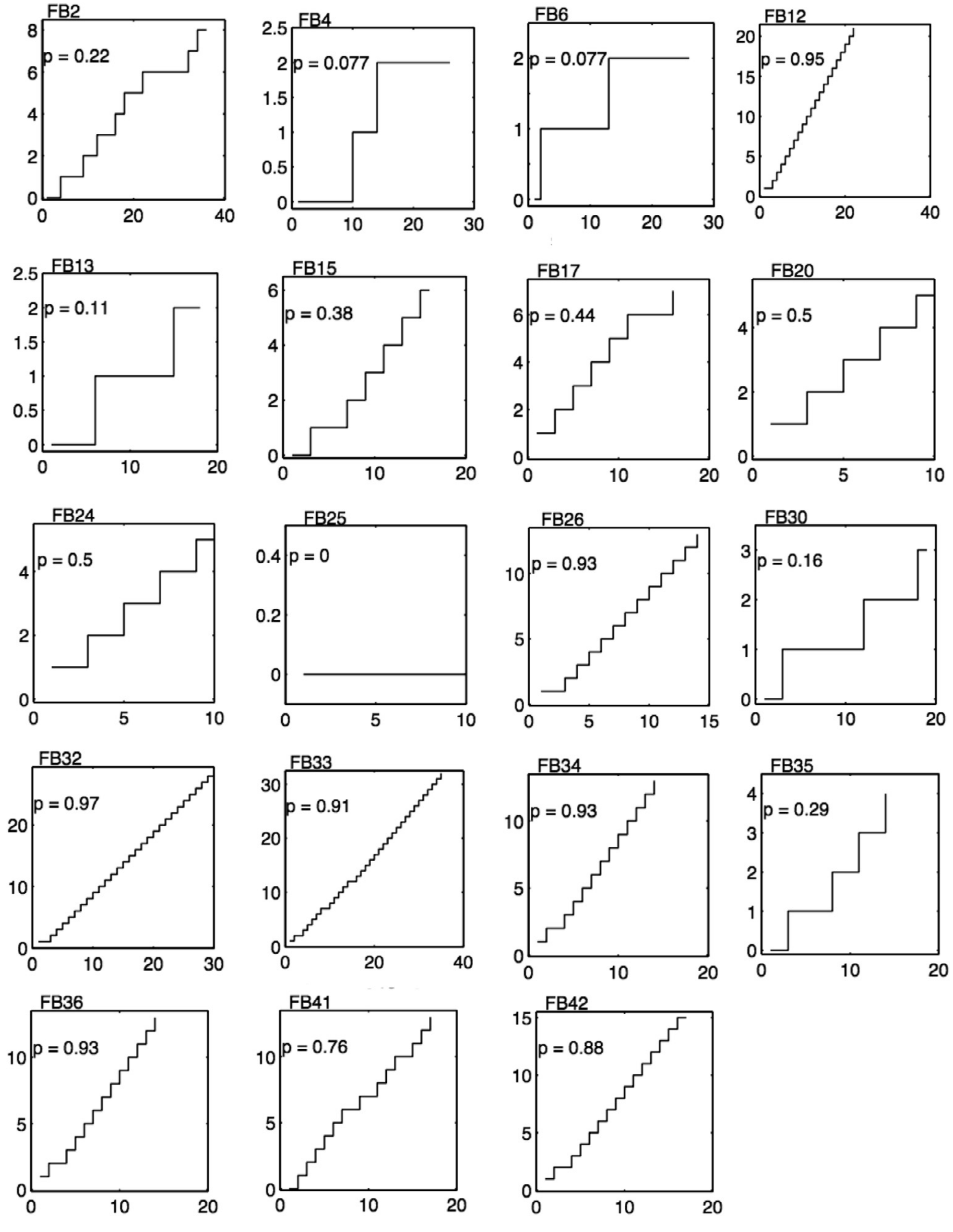

Fig. 8. Weak evidence for a model of no change (n = 11 US; n = 8 UK). X-axis: number of trials; Y-axis: number of successes.

4.2.1. Sudden change

A handful of records reflected a pattern of 'sudden insight' on the child's part, where performance changed dramatically from floor to ceiling. However other patterns of 'sudden' change were also observed, which did not demonstrate 'sudden insight'. When 'sudden' changes do occur they are often

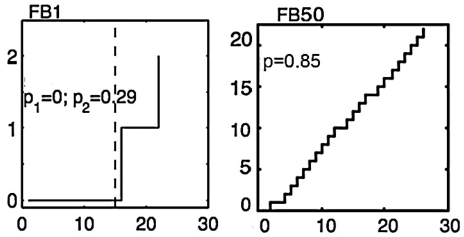

Fig. 9. Neutral (uninformative) evidence with respect to models of change or no change (n = 1 US; n = 1 UK). X-axis: number of trials; Y-axis: number of successes.

Table 3
Percentage of individual records showing evidence for change, no change, or uninformative evidence across US and UK cohorts.

Substantial evidence for		Weak evidence for		Uninformative
Change	No change	Change	No change	
21%	9%	21%	44%	5%

to asymptotic passing, but they can also occur to much lower levels, or from an intermediate level to ceiling performance.

4.2.2. Vacillations

A striking finding here is that, regardless of whether change was observed at any point throughout our period of observation, many children showed ‘stably unstable’ performance (mean probability of passing somewhere between 10% and 90% over several months with no significant change). Many of our participants went through a prolonged period of unstable performance, sometimes failing and sometimes passing. These fluctuations are reminiscent of findings from the animal learning literature. When Gallistel and colleagues used a similar change-point analysis of individual behavior in conditioned learning in animals, they found that onset of the conditioned response was sudden, and not gradual, with great individual variability in both latency of onset and in the asymptotic levels to which responding rose. The traditional smooth learning curve is thus simply an artifact of group averaging (Gallistel et al., 2004; Papachristos & Gallistel, 2006).

Both intra- and inter-individual variability in long-term performance has also been found in children’s theory of mind performance. Flynn et al.’s (2004) microgenetic study showed that children’s performance during a transitional period of development is likely to progress non-monotonically. Flynn (2006) tested preschool children repeatedly for five months on several theory-of-mind tasks. The majority of children’s performance varied from one month to the next. Thirty-seven percent of changes from one session to the next were improvements, while 24% of changes were regressions. In motor development, Adolph et al. (2008) have shown that changing the sampling of performance from monthly to daily measurements drastically altered the number of records showing variability, with only 9% of monthly records showing variability while 84% of daily records were variable.

These ‘stably unstable’ records, along with the ‘sudden change’ records that did not show ‘sudden insight’, illustrate the fact that performance at the level of an individual child is not binary and we cannot always categorize a child in terms of “failing” or “passing”. At the same time, we know from other work that individual differences in theory of mind can be stable over time (Wellman et al., 2008). These individual differences predict other important abilities in a child’s development, such as their responsiveness to feedback from teachers in a school setting (Lecce, Caputi, & Hughes, 2011). One challenge that the present findings pose is how we explain the apparent stability, as well as the variability, of individual differences over time. Methods such as the Bayesian change point analysis of individual records can help in this endeavor.

4.2.3. Decreases

In addition to non-monotonic increases in the probability of passing we observed two records of false belief attribution where the probability of passing actually decreased, although the evidence for real change in those records was weak in both cases (FB52 and FB43). In any event, such findings highlight the importance of examining individual records with methods that allow us to measure small, but real, changes in performance. Discovering and describing the inter-individual variability that exists can only serve to enhance our understanding of the mechanisms driving developmental change. Knowing that even one child's developmental profile follows a certain pattern tells us that the pattern is not impossible, and such a fact has implications for the constraints we build into our theories of developmental change. We made the decision to continue to test even the initial passers, who were at ceiling at the beginning of their participation, so as not to pass up the opportunity to observe rare regressions. It is noteworthy that nearly all initial passers continued to demonstrate high levels of passing throughout the entire testing period. However, one child (FB52) presented with a regression in performance, from 91% to 67%. Even if such occurrences are rare, with future studies, the slow accumulation of even rarities can be quantified—another advantage of multiple single-case studies.

Our findings show that variability is common and can last for periods up to a year or more! The degree of individual variability in asymptotic passing levels on the false belief task remains to be determined, along with what 'passing' really means for individual children. Answering this important question will require quantitatively analyzing individual records gathered over very long periods of time.

4.3. Some issues arising from the study of variability in development

Despite our attempts to vary task details, repeatedly testing the same child on the same underlying task may have altered that child's performance profile. However, we never provided feedback. Also false belief and true belief "situations" occur naturally at some unknown rate in the environment outside of testing and our testing sessions may not have added significantly to those learning opportunities. Moreover, we replicated our method and obtained strikingly similar results in the New Jersey and Bristol cohorts. Moreover, the coarse grain of our data conforms to the general pattern that has emerged over numerous cross-sectional studies (see Wellman et al.'s (2001) meta-analysis). Our participants were approximately equally likely to pass or fail around their fourth birthday with lower and higher probabilities of passing at lower and higher ages, respectively, showing the classic pattern. Finally, the five children we tested on the true belief control task showed reassuringly consistent high levels of passing throughout. All of these features lead us to believe that the 'instability' we observed is real.

4.3.1. Premises of the Bayesian change point algorithm

There is no assumption of linearity underlying the Bayesian change point analysis, making it appropriate for modeling any developmental process (similar to non-linear growth curves; Grimm et al., 2011; Van Geert, 1998). Indeed several of our participants showed a 'sudden' developmental change (which we distinguish from 'sudden insight', moving swiftly from consistently failing to consistently passing a task). Gallistel et al. (2004) rigorously defined the question of what 'sudden change' means with respect to conditioned learning in animals. In the present study, 'sudden change' simply means on or between the trial indicated and the previous trial. There is no assumption of fixed intervals between responses underlying the Bayesian change point analysis. Because the interval between trials varied in our exploratory study (with an upper limit of 6 weeks), 'sudden' has no precise chronological meaning. The length or variability in length of inter-session periods does not affect the statistical test, but the reader should be aware that 'sudden' is relative to session and not time (for a discussion of time as a variable in developmental research, see Ram, Gerstorf, Fauth, Zarit, & Malmberg, 2010). Mapping model parameters to developmental processes is a crucial consideration (Ferrer & McArdle, 2010). When developmental change does occur in an individual child, does it occur instantaneously from one 'processing cycle' to the next or gradually over a period of seconds, minutes, hours, days, ...? No doubt, this is a very difficult question to probe experimentally in children; but, for us, the exciting thing is that Bayesian change-point analysis begins to make exploration of this question possible.

4.3.2. Evaluating the strength of the evidence for each model of the cumulative records

Only two cases presented nearly neutral evidence with respect to potential models of the data. The rest of the cases could be interpreted as conforming either to a model of change or a model of no change. The strength of the evidence garnered in each individual case is the primary currency of Bayesian change point analysis. As such, the factors affecting it deserve particular attention. We found more records with substantial evidence for change than with substantial evidence for no change (akin to the null hypothesis in traditional terms). It is harder to get substantial evidence for the null, rather than for the alternative to the null, because small samples and "noise" limit the possible strength of the evidence for the null, but not the possible strength of the evidence against it. We now discuss the reasons why this is the case.

Consider first the effect of sample size. Suppose we want to compare the relative likelihoods of two hypotheses about a coin: (1) It's a fair coin (the null); (2) It's a biased coin with a true probability of heads that is somewhere between .5 and 1 (the alternative to the null). We flip it 10 times and get 5 heads. This is, of course, exactly what the first hypothesis predicts. But how strongly does this result favor the first hypothesis over the second? Not very strongly because this result is also highly likely if the true p is .6 (in which case, of course, the second hypothesis is correct, because .6 is somewhere between .5 and 1). Thus, even when one gets *exactly* the outcome predicted by the null, the evidence in favor of the null cannot be very strong if the sample size is small. Contrast this with the case where we get 10 heads in 10 flips. The evidence against the null is now very strong, even though the sample size remains 10, because this outcome is very unlikely if $p = .5$ but very likely if $p > .9$ (another case in which the alternative hypothesis is true).

A similar point arises with regard to variability. With samples of modest size (say, between 5 and 25), it is possible to obtain very strong evidence against the null but not in favor of it, under reasonable assumptions about the alternative. Suppose that we are measuring, say, heights and our sample size is 10 French men and 10 English men and that the means of the two samples are identical (which, of course, almost never happens). Again, this is exactly what the null hypothesis – that the average height of French and English men is the same – predicts. Suppose that the alternative hypothesis is that the average height of French men is greater than that of English men by possibly as much as 5 cm. Now suppose that the pooled within-sample variance is 1 mm. In other words, all 20 men have the same height to within less than 1 mm! Under these low-variance conditions, the evidence would strongly favor the null. Now, suppose, more realistically, that the pooled within-sample variance is on the order of 25 cm. With the small sample size and the large variance, the evidence in favor of the null cannot be considered strong. But the evidence against the null can be. Suppose that the mean for the English men is 170 cm, whereas that for the French is 190. A difference in sample means this big is *much* more likely if the true difference is 10 cm (upper limit imposed by our formulation of the alternative) than if the true difference is 0.

The extent to which the length of the baseline and the intensity of data sampling will affect the evidence in support of a model remains an empirical question. The answer may well depend on the domain under investigation, with differing implications for, say, motor or cognitive development (Adolph et al., 2008; see also Van Geert & Van Dijk, 2002). In theory of mind, using the explicit false belief attribution task, sampling more frequently than once per month seems desirable, and starting the sampling before the third birthday would seem prudent. These *desiderata* will of course be balanced by practical considerations, but at least we begin to outline the characteristics of a study that could shed further light on the shape and rate of preschoolers' cognitive development at the individual level.

4.4. Future work using Bayesian change-point analysis

The change-point algorithm not only gives us a statistically principled way to determine whether or not an individual's cumulative record contains evidence of a change in performance, it also locates the best estimate of the point of change in that record. Thus, we can pinpoint when in the sequence of trials for that individual change is most likely to have taken place. Although in this initial study, we were not able to exploit this feature, in future it would be possible to study more than one type of task and to relate change points in multiple task records within an individual participant to establish lead-lag relationships (Ferrer & McArdle, 2010). This would give a new way to examine precursors and inter-task correlations across time at the level of the individual. For example, theoretical models that advocate that competences must follow a particular developmental sequence before later competences emerge, would benefit greatly from change-point analysis that establishes performance at the individual level. Here we have studied natural development in a task that is not formally instructed. But there are a number of other contexts, for example, educational and clinical treatment studies, where change point analysis of individual subjects linked to various pedagogical or therapeutic interventions could be crucial.

4.5. On modeling cognitive developmental processes like theory of mind

Our principal aim here was not to contrast the different theoretical approaches to theory of mind development but simply to illustrate how the Bayesian change-point analysis might be used through an initial study with children. However, it does seem to us that one upshot of these data is that the traditional notion of passing and failing the false belief task (and perhaps other developmental tasks) has limited value. Rather we suggest it is better to think in terms of the probability of passing: what changes with development is the probability of passing the false belief task, and it changes in different ways in different children at different times. Perhaps this fits better with a complex of performance systems whose interactions fluctuate daily but which become gradually better integrated, rather than with a single change in an underlying competence. From a practical point of view, a single trial recorded as a binary response does not reveal much about underlying states. The aggregation of such measures informs us only about the 'aggregated child'; unfortunately, it is not the aggregated child who represents and processes information and it is not the aggregated child who develops. In this vein, we gain new insight into the longstanding notion of the "chance performance" level in the false belief task (e.g., Devine & Hughes, 2014; Wellman et al., 2001; Wimmer & Perner, 1983; see also Carpenter et al., 2002 for insightful discussion). It appears that typically children develop through a range of probabilities, including those usually described as "chance levels" (see Fig. 3). Indeed, there may be a case for regarding the middle probability range as a distinguishable state (see individual curves for FB change subjects 21, 44, 28, 7, 29, and 43, and for FB no change subjects 15, 17, 20, and 24). In a study of toddlers' and adults' anticipatory eye-gaze in implicit tasks, Wang and Leslie (2016) show three distinct patterns of response. Perhaps it is time to stop regarding intermediate probabilities as without definite cause, or "chance", and start treating them as studiable in their own right.

One picture of theory of mind development that has been popular is "sudden insight" or some other one-time change from a conceptual-deficit state to a conceptually competent state. A variety of routes for such change have been proposed (e.g., Apperly & Butterfill, 2009; Carlson & Moses, 2001; Gopnik & Wellman, 2012; Perner, 1991; Perner & Roessler, 2012; Wellman et al., 2001). We found such dramatic changes in individual theory of mind only in a small minority (around 12%) of preschoolers. These children indeed showed a sudden shift from floor to ceiling performance between one testing session and another. Clearly, such cases are of interest even if not typical, as are those even rarer individuals who show regression in performance.

Instead, we believe that our findings fit most straightforwardly with a performance systems account in which sufficient processing resources (e.g., inhibitory executive function) are required to express underlying competence (e.g., Leslie, German, & Polizzi, 2005; Leslie & Polizzi, 1998; see also Mahy et al., 2014). The present results show that, whereas these resources will tend to increase over the very long term, they can fluctuate at shorter time scales. Thus, performance typically shows neither a gradual and steady improvement nor sudden insight. Such instability in performance can occur at varying levels and persist over fairly long periods, for example, up to a year or more. More recent versions of "theory-theory" (Gopnik & Wellman, 2012) have adopted a Bayesian computational account, in which the child randomly accesses and tests multiple hypotheses about belief, gradually revising and replacing their initial hypotheses. The move to a computational account is welcome, but each hypothesis in such a model adds an additional parameter, allowing it to fit more complex curves only at the expense of increasing model complexity. A challenge for future modeling is to optimize fit while minimizing complexity.

It may be that none of the existing theories of theory of mind development are actually supported by our series of single case studies, simply because none of the current approaches actually make unqualified predictions about the shape, rate and variety of developmental change in the individual. Existing theories instead aim to characterize the starting or later 'steady' states in the child's cognitive system but largely leave open the question of what the transitions between these states look like. We do not think this is necessarily a bad thing because these two questions may even be somewhat orthogonal. The method we describe here gives for the first time a glimpse of the shape(s) of developmental change as this occurs in the individual, and provides the first data for quantitative models of these processes (Greenwald, 2011).

Finally, the methods we describe here are not limited to studies of developmental change but are applicable more widely to situations where changes in individuals undergoing extended therapies, experimental interventions, or education programs are the center of interest. The researcher, clinician, or pedagogue will be able to determine on an individual by individual basis whether or not the intervention actually produced change in behavior and, if so, when in the time series that change occurred.

Acknowledgments

Support for these studies was provided by NSF grants BCS-0725169 and BCS-0922184 to AML. Support was also provided to the first author by an ESRC grant to BMH. We would like to thank Nora Newcombe and the anonymous reviewers who provided useful comments on the manuscript.

Appendix A

Further technical information about the Bayesian change point analysis

What Bayesian model selection does that NHST does not do is adjudicate the trade-off between adequacy and simplicity among alternative descriptions of the data (Gallistel, 2009; MacKay, 2003). If we assume that a binary sequence was generated by a Bernoulli process with an unchanging p parameter, then we can describe it adequately simply by estimating that p. If we assume that it was generated by a Bernoulli process whose p parameter made a step change at some point during the sequence, then we can describe it adequately by estimating three parameters, the p before the change, the p after the change, and the change point, cp. If we assume that it was generated by a Bernoulli process whose p parameter changed twice, then we can adequately describe it by estimating 5 parameters: p_1, p_2, p_3, cp_1, and cp_2, and so on. The challenge is to decide what are the best descriptions of our experimentally obtained binary sequences, $\mathbf{D} = [d_1, d_2, \ldots, d_n]$. A more complex description, one involving more parameter estimates, will always describe the data more precisely than a simpler one, as may be seen by considering the limiting case in which there are $n - 1$ change points in the description of a sequence of length n. Estimating the p values is particularly simple in this case, because the estimate of p_i is 1 just in case d_i is 1 and the estimate of p_i is 0 just in case d_i is 0. The problem with this "model" (or summary description) of the data is that it does not summarize; it uses more parameters to describe the data than there are data to describe.

In Bayesian model selection, one chooses the model (that is, the summary of the data) with the greatest marginal likelihood. The marginal likelihood of any summary description is the integral of the product of the prior probability distribution for that description and the likelihood function. Among the charms of the basic Bayesian computation is that the marginal likelihood of a description is proportional to how well it describes the data and inversely proportional to an exponential function of its complexity, where its complexity is measured by the number of parameters that must be estimated from the data if one adopts that description. There is no subjectivity in this; the beliefs of the experimenter are irrelevant. The preference accorded a simpler but adequate description follows directly from the fact that a possible description of the data is represented by a prior distribution whose dimensionality equals the number of parameters to be estimated. This prior distribution specifies the uncertainty about the values of those parameters before the data are examined. Regardless of its dimensionality, a prior distribution always contains a unit mass of prior probability. The greater the dimensionality of the parameter space that supports the distribution, the more diffusely that unit mass must be spread around, because the volume of the parameter space in which it must be spread grows exponentially with the dimensionality of that space. A diffuse but widely spread prior probability density is an advantage when the data are such that only a more complex description can do them justice, but it is a handicap when the data allow of a simpler description that is as good or almost as good. In that case, much of the prior probability for the more complex description is out in portions of the parameter space where the likelihood function is essentially 0. When the likelihood function multiplies the prior probability distribution, that prior probability is nullified, resulting in a reduced marginal likelihood for the more complex description. The marginal likelihood measure favors the description that puts more prior probability where there is high likelihood (see Gallistel, 2009 for graphic illustration and further explanation).

Appendix B

Sample scenarios for the false belief tasks

B.1. False belief-unexpected transfer

The direction of transfer of the object, left to right or right to left, was counterbalanced. There were always three possible locations so that there were two ‘incorrect’ options on every trial.

> “Steve and Corrie are playing here in their room. In their room there is a basket and atoybox. Now Steve is going for a walk. He leaves his ball here in the basket and goes outside. He can’t see us anymore. While Steve is outside, look what happens! Corrie takes the ball from here, and he puts it in here. Now Steve didn’t see that happen, did he? No. Well, Steve is coming back soon, and I have some questions for you.”
> “Where is the ball now?” (Reality control question)
> “Where did Steve put the ball, in the beginning of the story?” (Memory control question)
> “Where does Steve think the ball is?” (Think test question)
> “Does Steve know where the ball is?” (Know test question)

Other scenarios included, e.g., Dino leaves a fish in one of three coloured boxes while going to school; Annabelle leaves a plant in one of three pieces of furniture while riding her bike.

B.2. False belief-unexpected contents

> “I have something I want to show you.” (Show a crayon box)
> “What do you think is inside?” (control question)
> (Open box to reveal unexpected contents – a sock)
> “Now, if we show this box to Billy in your class, what will Billy think is inside?”
> (Third person think test question)
> “When I first showed it to you, what did you think was inside?” (First person think test question)

Other scenarios included, e.g., a spoon in a soup box; a small toy cow in an egg carton.

Appendix C

Detailed individual data is presented in the table that follows: age at first test (in months), profile number corresponding to the individual cumulative records of performance, and Bayes factor corresponding to each cumulative record, by cohort. Note that BF’s for initial passers are represented in bold; also shown is Weight of Evidence ($\log_{10}$(BF)), positive values support Change, negative support No Change.

	Age at first test	Profile	Bayes factor	Weight of evidence
US cohort (1)				
	34	FB01	1.01	0.00
	36	FB02	0.47	−0.33
	39	FB03	0.25	−0.60
	39	FB04	0.39	−0.41
	40	FB05	1459.70	3.16
	42	FB06	0.56	−0.25
	42	FB07	2.12	0.33
	46	FB08	0.28	−0.55
	46	FB09	2.83	0.45
	47	FB10	0.21	−0.68
	47	FB11	0.26	−0.59
	47	FB12	0.73	−0.14
	48	FB13	0.49	−0.31
	49	FB14	1811.50	3.26
	50	FB15	0.78	−0.11
	51	FB16	6.80	0.83
	51	FB17	0.84	−0.08
	51	**FB18**	**0.41**	−**0.39**
	51	**FB19**	**0.41**	−**0.39**
	52	FB20	0.84	−0.08
	53	FB21	9.32	0.97
	54	FB22	4.63	0.67
	55	**FB23**	**0.86**	−**0.07**
	56	FB24	0.84	−0.08
	58	FB25	0.45	−0.35
	58	FB26	0.79	−0.10
	62	**FB27**	**0.53**	−**0.28**
Mean	48.30		122.53	
sd	7.00		438.84	
UK cohort (2)				
	36	FB28	4.44	0.65
	38	FB29	2.22	0.35
	40	FB30	0.60	−0.22
	40	FB31	10679.00	4.03
	42	FB32	0.69	−0.16
	42	FB33	0.59	−0.23
	43	FB34	0.63	−0.20
	43	FB35	0.83	−0.08
	44	FB36	0.63	−0.20
	44	**FB37**	**0.44**	−**0.36**
	44	FB38	1.29	0.11
	45	FB39	4.88	0.69
	45	FB40	1.76	0.25
	46	FB41	0.71	−0.15
	47	FB42	0.89	−0.05
	48	FB43	1.55	0.19
	48	FB44	3.59	0.56

Appendix C (*continued*)

	Age at first test	Profile	Bayes factor	Weight of evidence
	49	FB45	1.26	0.10
	51	**FB46**	**0.27**	**−0.57**
	51	**FB47**	**0.30**	**−0.52**
	53	FB48	1.17	0.07
	54	FB49	1.26	0.10
	54	FB50	0.99	0.00
	54	**FB51**	**0.67**	**−0.17**
	55	**FB52**	**1.01**	0.00
Mean	46.24		428.47	
sd	5.36		2135.53	

References

Adolph, K. E., Robinson, S. R., Young, J. W., & Gill-Alvarez, F. (2008). What is the shape of developmental change? *Psychological Review, 115*, 527–543.

Agresti, A. (2000). Random-effects modeling of categorical response data. *Sociological Methodology, 30*(1), 27–80.

Amsterlaw, J., & Wellman, H. M. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development, 7*, 139–172.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116*, 953–970.

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

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14*, 110–118.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*, 37–46.

Bonawitz, E., Van Schijndel, T. J. P., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation and learning. *Cognitive Psychology, 64*, 215–234.

Boom, J., Wouters, H., & Keller, M. (2007). A cross-cultural validation of stage development: A Rasch re-analysis of longitudinal socio-moral reasoning data. *Cognitive Development, 22*, 213–229.

Bowers, J., & Davis, C. (2011). More varieties of Bayesian theories, but no enlightenment. *Behavioral and Brain Sciences, 34*, 193–194.

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis patterns of impaired performance: The case for single-patient studies. *Brain and Cognition, 5*, 41–66.

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

Carpenter, M., Call, J., & Tomasello, M. (2002). A new false belief test for 36-month-olds. *British Journal of Developmental Psychology, 20*, 393–420.

Cheshire, A., Muldoon, K. P., Francis, B., Lewis, C. N., & Ball, L. J. (2007). Modelling change: New opportunities in the analysis of microgenetic data. *Infant and Child Development, 16*, 119–134.

Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition, 23*(1), 43–71.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development, 9*, 377–395.

Cohen, A., & German, T. (2009). Encoding of others' beliefs without overt instruction. *Cognition, 111*, 356–363.

Devine, R. T., & Hughes, C. (2013). Silent films and strange stories: Theory of mind, gender, and social experiences in middle childhood. *Child Development, 84*, 989–1003.

Devine, R. T., & Hughes, C. (2014). Relations between false belief understanding and executive function in early childhood: A meta-analysis. *Child Development, 85*, 1777–1794.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives in Psychological Science, 6*, 274–290.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*, 134–140.

Ferrer, E., & McArdle, J. J. (2010). Longitudinal modeling of changes in psychological research. *Current Directions in Psychological Science, 19*, 149–154.

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

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*, 103–115.

Friedman, O., & Leslie, A. M. (2004a). A developmental shift in processes underlying successful belief-desire reasoning. *Cognitive Science, 28*, 963–977.

Friedman, O., & Leslie, A. M. (2004b). Mechanisms of belief-desire reasoning: Inhibition and bias. *Psychological Science, 15*, 547–552.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116*, 439–453.

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Science, 101*, 13124–13131.

Glymour, C. (1994). On methods of cognitive neuropsychology. *British Journal for the Philosophy of Science, 45*, 815–835.

Goldin-Meadow, S., Alibali, M. W., & Church, R. B. (1993). Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review, 100*, 279–297.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review, 118*, 110–119.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory. *Psychological Bulletin, 138*(6), 1085.

Greenwald, A. G. (2011). There is nothing so theoretical as a good method. *Perspectives on Psychological Science, 7*, 99–108.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science, 21*, 263–268.

Grimm, K. J., Ram, N., & Hamagami, F. (2011). Nonlinear growth curves in developmental research. *Child Development, 82*, 1357–1371.

Hoffman, D. A., Jacobs, R., & Gerras, S. J. (1992). Mapping individual performance over time. *Journal of Applied Psychology, 77*, 185–195.

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

Hood, B. M., Cole-Davies, V., & Dias, M. (2003). Looking and search measures of object knowledge in preschool children. *Developmental Psychology, 39*, 61–70.

Howe, C., Tavares, J. T., & Devine, A. (2012). Everyday conceptions of object fall: Explicit and tacit understanding during middle childhood. *Journal of Experimental Child Psychology, 111*(3), 351–366.

Howe, C., Taylor Tavares, J., & Devine, A. (2014). Children's conceptions of physical events: Explicit and tacit understanding of horizontal motion. *British Journal of Developmental Psychology, 32*(2), 141–162.

Hughes, C., Adlam, A., Happé, F., Jackson, J., Taylor, A., & Caspi, A. (2000). Good test-retest reliability for standard and advanced false belief tasks across a wide range of abilities. *Journal of Child Psychology and Psychiatry, 41*, 483–490.

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

Jeffreys, H. (1961). *The Theory of Probability*. OUP Oxford.

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). A longitudinal study of mathematical competencies in children with specific mathematics difficulties versus children with comorbid mathematics and reading difficulties. *Child Development, 74*, 834–850.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversations: The role of mutual knowledge in comprehension. *Psychological Science, 11*, 32–38.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14*, 293–300.

Lamb, R. L., Vallett, D., & Annetta, L. (2014). Development of a short-form measure of science and technology self-efficacy using Rasch analysis. *Journal of Science Education and Technology, 23*(5), 641–657.

Lecce, S., Caputi, M., & Hughes, C. (2011). Does sensitivity to criticism mediate the relationship between theory of mind and academic achievement? *Journal of Experimental Child Psychology, 110*(3), 313–331.

Lee, M. D., & Wagenmakers, E. J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review, 112*, 662–668.

Leslie, A. M., German, T., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology, 50*, 45–85.

Leslie, A. M., & Polizzi, P. (1998). Inhibitory processing in the false belief task: Two conjectures. *Developmental Science, 1*, 247–254.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. New York: Cambridge University Press.

Mahy, C. E. V., Moses, L. J., & Pfeifer, J. H. (2014). How and where: Theory-of-mind in the brain. *Developmental Cognitive Neuroscience, 9*, 68–81.

Math Works. (2000). MatLab. The MathWorks Inc., Natick, MA, 2000.

Mayes, L., Klin, A., Tercyak, K. P., Cicchetti, D. V., & Cohen, D. J. (1996). Test-retest reliability for false belief tasks. *Journal of Child Psychology and Psychiatry, 37*, 313–319.

Nesselroad, J. R., Gerstorf, D., Hardy, S. A., & Ram, N. (2007). Idiographic filters for psychological constructs. *Measurement, 5*, 217–235.

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

Opfer, J. E., & Siegler, R. S. (2004). Revisiting preschoolers' *living things* concept: Microgenetic analysis of conceptual change in basic biology. *Cognitive Psychology, 49*, 301–332.

Papachristos, S., & Gallistel, C. R. (2006). Autoshaped head poking in the mouse: A quantitative analysis of the learning curve. *Journal of the Experimental Analysis of Behavior, 85*, 293–308.

Pears, K. C., & Moses, L. J. (2003). Demographics, parenting, and theory of mind in preschool children. *Social Development, 12*, 1–20.

Perner, J. (1991). *Understanding the Representational Mind*. The MIT Press.

Perner, J., & Roessler, J. (2012). From infants' to children's appreciation of belief. *Trends in Cognitive Science, 16*, 519–525.

Piaget, J. (1954). *The construction of reality in the child*. New York: Basic Books.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(04), 515–526.

Ram, N., Gerstorf, D., Fauth, E., Zarit, S., & Malmberg, B. (2010). Aging, disablement, and dying: Using time-as-process and time-as-resources metrics to chart late-life changes. *Research in Human Development, 7*, 27–44.

Ruffman, T., Garnham, W., Import, A., & Connolly, D. (2001). Does eye gaze indicate implicit knowledge of false belief? Charting transitions in knowledge. *Journal of Experimental Child Psychology, 80*, 201–224.

Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., & Lee, K. (2006). The development of executive functioning and theory of mind: A comparison of Chinese and U.S. preschoolers. *Psychological Science, 17*, 74–81.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. New York: Holt.

Shallice, T., & Evans, M. E. (1978). The involvement of the frontal lobes in cognitive estimation. *Cortex, 14*, 294–303.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116*, 250–264.

Siegler, R. S. (2007). Cognitive variability. *Developmental Science, 10*, 104–109.

Siegler, R. S., & Svetina, M. (2002). A microgenetic/cross-sectional study of matrix completion: Comparing short-term and long-term change. *Child Development, 73*, 793–809.

Singer, J. D., & Willett, J. B. (1991). Modeling the days of our lives: Using survival analysis when designing longitudinal studies of duration and the timing of events. *Psychological Bulletin, 110*, 268–290.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*, 1279–1285.

Van Dijk, M., & Van Geert, P. (2007). Wobbles, humps and sudden jumps: A case study of continuity, discontinuity and variability in early language development. *Infant and Child Development, 16*, 7–33.

Van Geert, P. (1998). We almost had a great future behind us: The contribution of non-linear dynamics to developmental-science-in-the-making. *Developmental Science, 1*, 143–159.

Van Geert, P., & Van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior and Development, 25*, 340–374.

Wang, L., & Leslie, A. M. (2016). Is implicit theory of mind the 'real deal'? The own-belief/true-belief default in adults and young preschoolers. *Mind & Language, 31*, 147–176.

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.

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

Wellman, H. M., Lopez-Duran, S., LaBounty, J., & Hamilton, B. (2008). Infant attention to intentional action predicts preschool theory of mind. *Developmental Psychology, 44*, 618–623.

Wetzels, R., Matzke, D., Lee, M. L., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in empirical psychology: An empirical comparison using 855 *t*-tests. *Perspectives in Psychological Science, 6*, 291–298.

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the design of studies of individual growth. *Educational and Psychological Measurement, 49*, 587–602.

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

Young, G. S., Rogers, S. J., Hutman, T., Rozga, A., Sigman, M., & Ozonoff, S. (2011). Imitation from 12 to 24 months in autism and typical development: A longitudinal Rasch analysis. *Developmental Psychology, 47*(6), 1565.

Zelazo, P. D., Frye, D., & Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cognitive Development, 11*, 37–63.