CORTEX 86 (2017) 205–215

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Special issue: Note

Does the way we read others' mind change over the lifespan? Insights from a massive web poll of cognitive skills from childhood to late adulthood

David Klindt [a], Marie Devaine [a] and Jean Daunizeau [a,b,*]

[a] Brain and Spine Institute (ICM), Paris, France
[b] Translational Neuromodelling Unit, ETH, Zurich, Switzerland

ARTICLE INFO

Article history:
Received 1 February 2016
Reviewed 2 March 2016
Revised 11 July 2016
Accepted 9 September 2016
Published online 23 September 2016

Keywords:
Theory of mind
Dual process theory
Cognitive control
Executive functions
Cognitive development
Cognitive decline

ABSTRACT

Mentalizing or Theory of Mind (ToM), i.e., the ability to recognize what people think or feel, is a crucial component of human social intelligence. It has been recently proposed that ToM can be decomposed into automatic and controlled neurocognitive components, where only the latter engage executive functions (e.g., working memory, inhibitory control and task switching). Critical here is the notion that such dual processes are expected to follow different developmental dynamics. In this work, we provide novel experimental evidence for this notion. We report data gathered from about thirty thousand participants of a massive web poll of people's cognitive skills, which included ToM and executive functions. We show that although the maturation of executive functions occurs in synchrony (around 20 years of age), this is not the case for different mentalizing competences, which either mature before (for elementary ToM constituents) or after (for higher-level ToM). In addition, we show that inter-individual differences in executive functions predict variability in higher-level ToM skills from the onset of adulthood onwards, i.e., after the complete maturation of executive functions. Taken together, these results indicate that the relative contribution of ToM's controlled component significantly changes with age. In particular, this implies that, over the lifespan, people may rely upon distinct cognitive architectures when reading others' minds.

1. Introduction

How do you know what others think or feel? Mentalizing or Theory of Mind (ToM), i.e., the ability to identify covert mental states from the interpretation of overt social signals (ranging from eye gazes and facial expressions to behavior and language), is a crucial component of human social intelligence (Frith & Frith, 2012). This is because ToM endows humans with highly adaptive social skills such as bonding, teaching or deceiving, whose sophistication is arguably unique within the animal kingdom (Call & Tomasello, 2008; Penn & Povinelli, 2007). But how stable is the cognitive architecture that

* *Corresponding author*. Motivation, Brain and Behaviour Group, Brain and Spine Institute, 47, bvd de l'Hopital, 75013, Paris, France.
E-mail address: jean.daunizeau@gmail.com (J. Daunizeau).
http://dx.doi.org/10.1016/j.cortex.2016.09.009

enables people to read others' mind over the lifespan? In particular: does the contribution of executive functions (e.g., working memory, inhibitory control, etc…) to mentalizing abilities change from childhood to late adulthood? These are the questions we address in this work.

Understanding others' mental states is a developing ability, whose most elementary constituents are acquired during early childhood (Wellman, Cross, & Watson, 2001). This development starts early, since children in their second year of life already show signs of surprise when others do not behave in accordance with their beliefs (Onishi & Baillargeon, 2005). This is taken as evidence of children's insight that people's behavior is driven by their beliefs rather than by physical reality, even if these beliefs happen to be false. ToM's sophistication then culminates at adulthood, when it engages a specific large-scale brain network, typically including the precuneus, the temporo-parietal junction and the medial prefrontal cortex (Gallagher & Frith, 2003; Van Overwalle & Vandekerckhove, 2013). This is not to say, however, that ToM is a monolithic cognitive ability. We know from dissociations observed in patients (e.g., autism spectrum disorder or Williams syndrome) that ToM can be decomposed into distinct cognitive subcomponents (Senju, Southgate, White, & Frith, 2009; Tager-Flusberg & Sullivan, 2000). We also know that mentalizing competences vary greatly between neurotypical adults and show weak inter-task correlations (Ferguson & Austin, 2010; Flobbe, Verbrugge, Hendriks, & Krämer, 2008; Lebreton, Kawa, d'Arc, Daunizeau, & Pessiglione, 2012). In fact, variations in the volume of elements of the ToM brain network predict inter-individual differences in distinct mentalizing tasks (Cullen, Kanai, Bahrami, & Rees, 2014; Hooker, Bruce, Lincoln, Fisher, & Vinogradov, 2011; Lewis, Rezaie, Brown, Roberts, & Dunbar, 2011). In addition, it has been shown that performances in various mentalizing tasks are correlated with measures of working memory and inhibitory control (Carlson & Moses, 2001; German & Hehman, 2006; Gordon & Olson, 1998). The contribution of such domain-general executive functions has been further evidenced by experimental studies demonstrating that some sophisticated mentalizing processes are disrupted by the concurrent engagement in secondary cognitively-demanding tasks (Apperly, Samson, & Humphreys, 2009; Bull, Phillips, & Conway, 2008; Lin, Keysar, & Epley, 2010; Qureshi, Apperly, & Samson, 2010). This multi-faceted portrait is compatible with a dual process theory of ToM (Frith & Frith, 2008). In brief, this theory suggests that full-grown mentalizing relies on both specialized representational skills (the ability to represent mental states as such) as well as executive resources for goal-oriented (i.e., task-related) processing of these representations (German & Hehman, 2006). Over the course of development, the representational system specializes for tracking mental states in an automatic, fast and efficient way. Its elementary constituents are expected to mature much before cognitive control, which enables the flexible allocation of executive resources. In this view, mind-reading is analogous to text-reading, in that an increasing part of its constituent cognitive processes (such as visual word recognition) become implicit and automatic as people grow older (Heyes & Frith, 2014). Would this idea hold true, it would imply that different ToM competences would be based upon qualitatively distinct cognitive architectures, whose relative contribution to mind reading may change with age.

In this work, we provide preliminary evidence that supports and extends this notion. We report data gathered from the BRAiN'US project, a free smartphone app that allows us to perform a massive web poll of some specific set of people's cognitive skills (https://sites.google.com/site/brainusapp2/). Here, we summarize the performance results in six games, which were designed to assess increasingly sophisticated mentalizing abilities (see below) and distinct executive functions (working memory, inhibitory control and task switching), respectively. In brief, we segmented our large sample into 16 age groups, ranging from 5 to 85 years old (age bin span = 5 years). We then quantified the lifespan dynamics of both mean performances and statistical interdependencies among these. The former allows to quantifying the time course of development and decline of investigated cognitive functions. The latter enable us to directly assess age-related changes in the contribution of executive functions to mentalizing abilities.

2. Methods

Recruitment of participants was performed through the smartphone/internet BRAiN'US platform (https://sites.google.com/site/brainusapp2/). This study was approved by a non-governmental ethics committee for academic research (CPP – Ile de France 1) on the 29th of July 2014, and was declared to the CNIL (i.e., the French national commission on informatics and liberties), under the name "massive web poll of the population's cognitive skills". Accordingly, participants were informed about the objectives and context of the project, and their consent was sought at the time of registration and then prior to engaging in each test. Data were then recorded on an anonymous and secure web database, along with biographical information including age, gender, place of residence, educational level and mental health status (under participants' conditional acceptance). All statistical data analyses were performed using the VBA freeware (Daunizeau, Adam, & Rigoux, 2014).

Subjects could play any of the BRAiN'US games in any order (although presentation order was randomized across subjects), and they could freely call off the experiment at any point. Before the beginning of each test, subjects were provided with written instructions accompanied with graphical summaries of the task. They then went through a training phase (which they could repeat as many times as they wanted). Feedback on their performance was provided at the end of each game. In this short note, we analyze performance data in the following six games:

- "*Emily and the donuts*" (*FB*): This is a variant of a false-belief task (1 trial), which evaluates one's ability to distinguish one's beliefs from others' beliefs (Wimmer & Perner, 1983). It can be seen as one of the most elementary constituent of ToM. Performance in this test is binary (correct *vs* incorrect answer).

- "*Triangles at the box-office*" (*anim*): This is a variant of the Frith-Happé animations test (White, Coniston, Rogers, & Frith, 2011), which evaluates one's ability to recognize others' intentions and emotions from their overt behavior. Performance was measured in terms of the rate of correct answers, when categorizing silent animations as one of the three following types: "no interaction", "physical interaction" and "mental interaction".
- "*Hide-and-seek*" (*HS*): This is a two-players competitive game (40 trials), in which players have to guess their opponent's next move. Here, participants play against on-line learning algorithms endowed with artificial ToM. This test measures one's ability to predict others' behavior in the context of strategic social interactions (Devaine, Hollard, & Daunizeau, 2014b, 2014a). Performance was measured in terms of the rate of correct answers, averaged across opponents.
- "*Three steps behind*" (*WM*): This is a variant of a 3-back test (128 trials), which evaluates one's working memory (Braver et al., 1997). Performance was measured in terms of the sensitivity index d-prime.
- "*Flyswatter*" (*gonogo*): This is a variant of a go-nogo test (250 trials), which evaluates one's inhibitory control (Aron, 2007). Performance was measured in terms of the sensitivity index d-prime.
- "*The perfect pair*" (*WCST*): This is a variant of the Wisconsin Card Sorting Test (48 trials), which evaluates one's flexibility in task switching (Nelson, 1976). Performance was measured in terms of the rate of correct answers.

Although crowdsourcing cognitive experiments enables researchers to reach massive sample sizes, they may suffer from reduced data quality, when compared to well-controlled laboratory conditions (Brown et al., 2014; Dance, 2015). Thus, we performed a few pre-processing checks to evaluate the data quality, and removed corrupted data when necessary. First, participants who declared a neurological or psychiatric disease were excluded from the analysis ($n = 1877$). We also excluded participants whose age was missing or nonsensical ($n = 258$). Second, we checked whether the data distribution was unimodal by deriving age-dependant empirical histograms of performances in each game (except *FB*). We then detected and excluded outliers (*WM*: $n = 27$; *gonogo*: $n = 30$; *anim*: $n = 3$). Third, due to software issues in the early phase of the BRAiN'US project, we dismissed all data that had been acquired before the 4th January 2015 (*WM*: $n = 28$; *gonogo*: $n = 51$; *anim*: $n = 27$; *FB*: $n = 54$).

Before presenting the main results of this work, let us now summarize the main characteristics of the resulting population sample. Most players only played a reduced subset of all games; Fig. 1 below summarizes the number of players per game and per age bin.

Fig. 1A shows how representative our subjects' sample is. In particular, our sample could be improved by increasing the number of young subjects (below 20 years-old). One can also see that each game is played by approximately one among four participants, irrespective of game or age. However, this is largely compensated by the otherwise unusual sample size we have access to (for example, even the youngest age group includes 202 participants). In addition, the study's drop-off rate can be assessed on Fig. 1B, in terms of the exponential decay of the number of participants who played a given number of games. In particular, one can see that most participants only engaged with a narrow subset of the available games (half-life is about 1.2 games). This provides a rough estimate of the proportion of missing data when assessing statistical dependencies between games' performances.

Some games were also played more than once by some participants. This allowed us to assess the test–retest reliability of performance scores, which is summarized on Fig. 2.

Except for *FB*, each game's test–retest reliability can be measured in terms of the correlation between performance scores during first and second game plays (across participants). In brief, we found a significant rank correlation (as measured using Kendall's tau) between first and second test administrations for each game (*anim*: $r = .66$, $\tau = .49$, $p < 10^{-6}$; *HS*: $r = .43$, $\tau = .18$, $p < 10^{-6}$; *WM*: $r = .11$, $\tau = .08$, $p = .04$; *gonogo*: $r = .81$, $\tau = .63$, $p < 10^{-6}$; *WCST*: $r = .38$, $\tau = .26$, $p < 10^{-6}$). Note that, in principle, *FB*'s test–retest reliability could be assessed using a chi-squared test. Although we failed to reject the null hypothesis (i.e., independent test administrations: $\chi^2 = .64$, $p = .42$), this provides no evidence against test–retest reliability in *FB*. We comment on this and related statistical issues in the discussion section.

3. Results

First, we estimated the effect of age on performance for each game. This is summarized on Fig. 3 below.

One can see that age has a profound impact on test performances, which is qualitatively similar across games. In summary, all test performances follow an inverted U-shaped pattern of cognitive development and decline, with an apex at the onset of early adulthood. More precisely, in all games except *HS*, performance at late adulthood (above 65 years-old) has fallen back to its early childhood's level (up until 8 years-old).

To quantify these effects, we regressed performance against a categorical model of development and decline, by partitioning participants into 16 distinct age groups (from 5 to 85 years old). We also included biographical information as secondary orthogonal factors in our analysis, i.e., our regression model was augmented with 16×2 additional regressors that encoded education level and gender (zscored within each age bin). This analysis allowed us to quantify the effect of education level and gender within each age group, above and beyond the main effect of age onto performance. Statistical significance of the age effect is confirmed by standard *F*-tests (*FB*: $F[15,8869] = 10.4$, $p < 10^{-6}$; *anim*: $F[15,6083] = 27.9$, $p < 10^{-6}$; *HS*: $F[15,5911] = 20.8$, $p < 10^{-6}$; *WM*: $F[15,4422] = 32.3$, $p < 10^{-6}$; *gonogo*: $F[15,8531] = 133.3$, $p < 10^{-6}$; *WCST*: $F[15,7516] = 47.7$, $p < 10^{-6}$). Nevertheless, the associated effect sizes are weak to moderate (*FB*: $R^2 = 1.7\%$; *anim*: $R^2 = 6.5\%$; *HS*: $R^2 = 5.0\%$; *WM*: $R^2 = 9.9\%$; *gonogo*: $R^2 = 19.0\%$; *WCST*: $R^2 = 8.7\%$). Interestingly, we also found that performance significantly increased with education level in all games except *HS*, although education explained a very weak amount of variance (*FB*: $F[1,8853] = 20.6$, $p = 6 \times 10^{-5}$, $R^2 = .2\%$; *anim*: $F[1,6067] = 153.9$, $p < 10^{-6}$, $R^2 = 2.5\%$; *HS*: $F[1,4955] = .1$, $p = .74$, $R^2 < .01\%$; *WM*: *F*

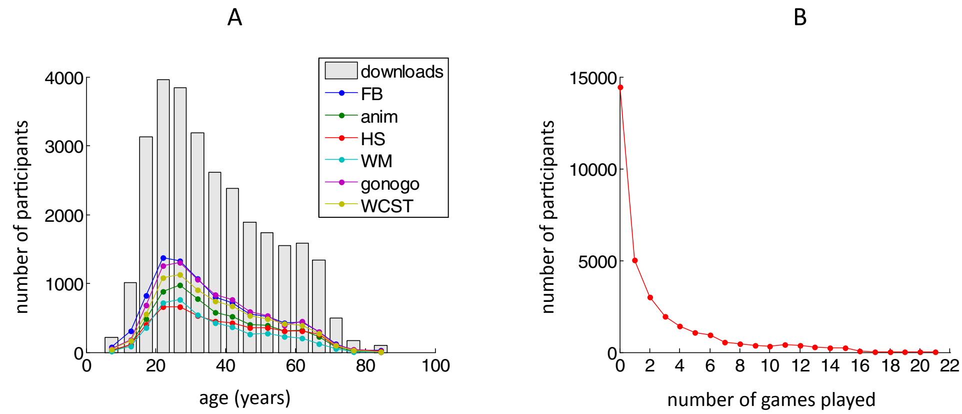

Fig. 1 – Basic characteristics of our sample. A: Number of participants (y-axis) as a function of age (x-axis), for each game (gray bars: app downloads, blue: *FB*, green: *anim*, red: *HS*, cyan: *WM*, violet: *gonogo*, yellow: *WCST*). Note that the bars are positioned at the mean age of the corresponding subgroup, which may not exactly match the arithmetic mean of the age range. B: Number of participants (y-axis) as a function of number of games played (x-axis).

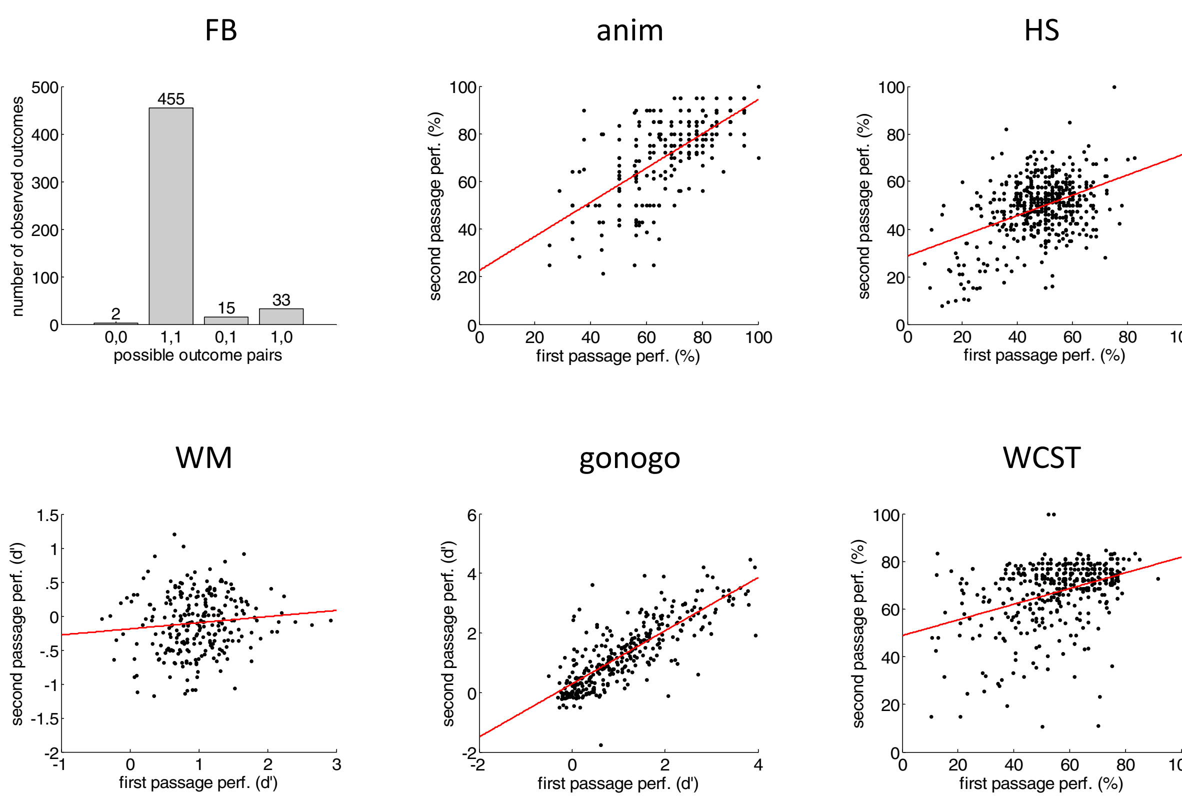

Fig. 2 – Test–retest reliability. In each panel except *FB* (top-left), performance during the second test passage (y-axis) is plotted against performance during the first passage (x-axis), for all participants who played the game twice. Red lines indicate the best-fitting affine transform of first-passage performances. Performances are expressed in their native measurement scale (see main text for details). For *FB*, we show the number of participants (y-axis) for each possible combination of test–retest outcomes (x-axis), where 0 = error and 1 = correct answer.

[1,4407] = 140.3, $p < 10^{-6}$, $R^2 = 3.1\%$; *gonogo*: $F[1,8515] = 71.4$, $p < 10^{-6}$, $R^2 = .8\%$; *WCST*: $F[1,7501] = 147.3$, $p < 10^{-6}$, $R^2 = 1.9\%$). We found no significant interaction effect between age and education level. We also found that gender had a significant effect in all executive games (but not social games), such that women perform slightly worse than men on average (*FB*: $F[1,8853]<.01$, $p = .98$; *anim*: $F[1,6067] = .2$, $p = .62$; *HS*: $F[1,4955] = 4.3$, $p = .04$; *WM*: $F[1,4407] = 26.5$, $p < 10^{-6}$; *gonogo*: $F[1,8515] = 153.9$, $p < 10^{-6}$; *WCST*: $F[1,7501] = 20.4$, $p = 6 \times 10^{-6}$). However, the effect size is very weak in all cases, i.e., gender explains very little performance variability (*FB*: $R^2<.01\%$; *anim*: $R^2<.01\%$; *HS*: $R^2 = .1\%$; *WM*: $R^2 = .6\%$; *gonogo*: $R^2 = 1.8\%$; *WCST*:

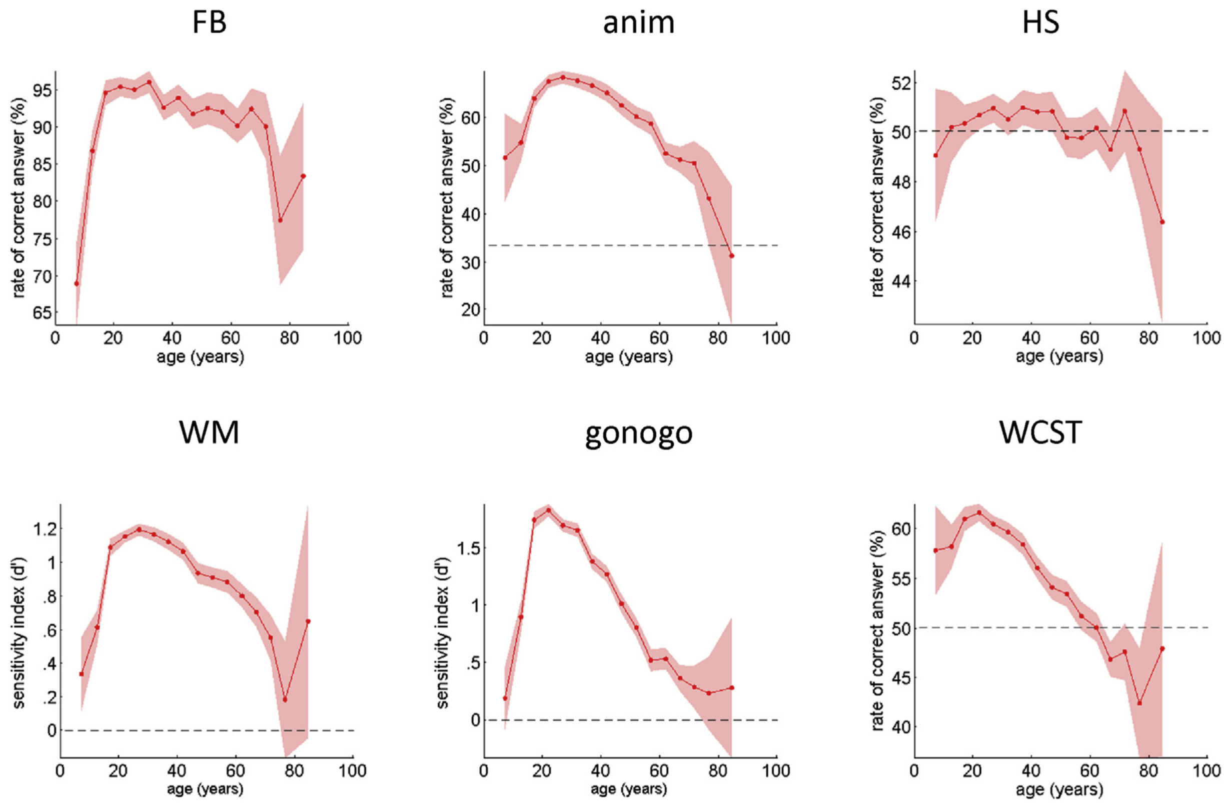

Fig. 3 – Cognitive development and decline. Mean performance (y-axis) is plotted as a function of age (x-axis) for each game (same as Fig. 2). Pink patches depict the 95% confidence intervals around the mean. Black dashed lines indicate chance level. Note that chance level is at 50% for *FB* (below the y-axis limits), and at 33.3% for *anim* because each trial consist of three alternative choices).

$R^2 = .3\%$). Here again, we found no significant interaction between age and gender.

Next, we asked whether the lifespan dynamics of cognitive development and decline is aligned across games. This is because social cognitive skills may rely upon different cognitive architectures, depending on whether their development is terminated before or after the maturation of executive functions. We first addressed the issue of estimating the age at which performance reaches its apex in each game. Thus, we evaluated the profile likelihood (Aitkin, 2014) of the performance apex, by fitting a piecewise-linear model with an unknown change point (apex) to performance data. The results are summarized on Fig. 4 below.

Fig. 4A depicts the fit accuracy of the piecewise-linear model for each game. One can see that the developmental and declining performance dynamics are well captured by the piecewise-linear model. This is despite its low degrees of freedom, which eventually result in moderate fit accuracy (*FB*: $R^2 = 1.7\%$; *anim*: $R^2 = 5.9\%$; *HS*: $R^2 = .3\%$; *WM*: $R^2 = 8.7\%$; *gonogo*: $R^2 = 18.8\%$; WCST: $R^2 = 8.0\%$). Fig. 4B shows the profile likelihood over candidate apices for each game. One can see that all games assessing executive functions are synchronous, i.e., their corresponding performance apex arise around very similar age ranges (*WM*: apex = 20.5 ± .7 years; *gonogo*: apex = 18.5 ± .4 years; *WCST*: apex = 20.2 ± 1.0 years; mean ± standard deviation). When testing for differences in apices across executive games (and correcting for multiple comparisons), we found no significant delay (*WM-gonogo*: $p = .02$; *WM-WCST*: $p = .85$; *WCST-gonogo*: $p = .11$). Games assessing social cognitive skills however, exhibit stronger differences. On the one hand, the performance apex of *FB* appears much sooner than that of executive functions (*FB*: apex = 15.1 ± .6 years). On the other hand, the performance apices of *anim* and *HS* are synchronous and are delayed w.r.t. executive functions (*anim*: apex = 27.5 ± 1.7 years; *HS*: apex = 29.7 ± 8.2 years). Among ToM games, we found that only *FB* and *anim* had significantly different apices (*FB-anim*: $p < 10^{-6}$; *FB-HS*: $p = .06$; *HS-anim*: $p = .74$). This is because the estimation uncertainty of *HS*'s performance apex is much wider than that of other games. In addition, all pairwise comparisons between the apices of ToM versus executive games were found significant, except for *HS*. These apex comparisons are summarized in Table 1 below.

Second, we asked whether performances evolve at different rates in these games. We thus focused on the absolute ratio between rates of decline and development, as measured using the two slopes of the fitted piecewise-linear model. By construction, this index is insensitive to the performance scale, which makes it directly comparable across different games. In what follows, we will refer to the decline/development rate ratio as the “relative decline rate”. Fig. 4C shows its Monte-Carlo sampling estimate for each game. Note that, as could have been expected from the apex analyses above, the estimation precision of *HS*'s

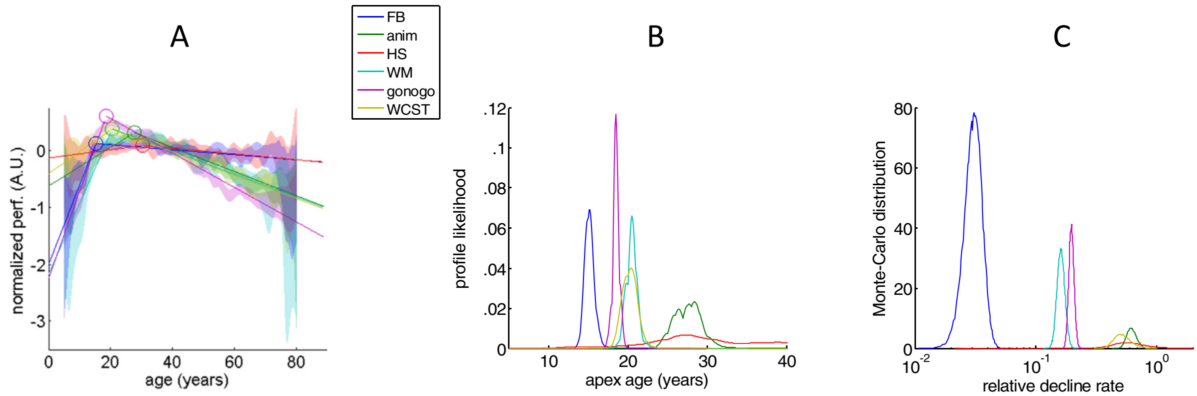

Fig. 4 – Piecewise-linear analysis of game performances. A: Colored plain lines indicate the best-fitting piecewise-linear model of zscored performance (y-axis) as a function of age (x-axis), for each game (same color code as in Fig. 1). Patches depicts the 95% confidence intervals of zscored performance for each game. B: The apex profile likelihood (y-axis) is plotted as a function of age (x-axis), for each game. B: Monte-Carlo probability distribution (y-axis) of the relative decline rate (x-axis, in log-scale), for each game. Note that the apparent identical spread of Monte-Carlo distributions is a side-effect of x-axis the log-scale.

Table 1 – Results of pairwise comparisons between estimated apices of ToM and executive games (in terms of the ensuing *p*-value).

	FB	*Anim*	*HS*
WM	$p < 10^{-6}$	$p = 1.8 \times 10^{-4}$	$p = .23$
gonogo	$p = 1.0 \times 10^{-5}$	$p < 10^{-6}$	$p = .15$
WCST	$p = 2.1 \times 10^{-5}$	$p = 2.7 \times 10^{-4}$	$p = .22$

relative decline rate is much lower than in other games. Nevertheless, one can see that all games exhibit a relative decline rate smaller than unity (*FB*: .03 ± .005; *anim*: .63 ± .06; *HS*: .75 ± .3; WM: .16 ± .01; *gonogo*: .20 ± .01; WCST: .55 ± .1; mean ± standard deviation). This means that performance rises more quickly during development than it decays during decline. In addition, here again, there is a dissociation between the games. To begin with, the relative decline rate for *FB* is significantly slower than in all other games except *HS* (all $p < 10^{-6}$, except *HS*: $p = .02$). Also, the relative decline rates in WM and *gonogo* are similar ($p = .03$), but significantly slower than in *anim* and WCST (all $p < 10^{-4}$). Lastly, quickly declining tasks show no significant difference in their relative decline rates (*HS-anim*: $p = .71$; *HS*-WCST: $p = .54$; WCST-*anim*: $p = .49$). All other comparisons are trivially significant (all $p < 10^{-6}$, except for *HS*).

So far, we have shown that mentalizing and executive skills can be dissociated based on performance apices and relative decline rates. In brief, cognitive development seems to follow a clear chronology, whereby low-level ToM matures earlier (and declines at a slower pace) than executive functions, which mature earlier (and decline at a slower pace) than high-level ToM. In what follows, we ask whether statistical dependencies between game performances are changing over the lifespan. This allows addressing more directly the question of whether the contribution of domain-general executive skills to mentalizing skills depends upon age. Fig. 5 shows the Kendall's pairwise correlation matrix between performance scores across participants within each age group.

One can see that, on average, there are only weak correlations between game performances. In fact, performance correlations seem to be highest across children, and, to a lesser extent, maybe also across old people. This is particularly salient for correlations among executive games, as well as between executive and ToM games. To quantify this, we first partitioned the pairwise correlation matrices into three subsets, namely: (i) correlations among ToM tasks, (ii) among executive tasks, and (iii) between executive and ToM tasks. Fig. 6A depicts the average Kendall's correlation in each subset as a function of age.

One can see that the temporal dynamics of all correlation subsets, except among ToM tasks, follow a decaying trajectory, with a peak at early childhood (ToM games: $\tau = -.07 \pm .11$; executive games: $\tau = .35 \pm .06$; ToM/executive games: $\tau = .13 \pm .05$; mean ± standard deviation). We found that the average correlations among executive tasks was significantly changing over the lifespan ($F[13,168] = 2.25$, $p = .01$, $R^2 = 14.8\%$). However, this was not the case for correlations among ToM tasks ($F[13,168] = 1.0$, $p = .44$, $R^2 = 7.2\%$), or for correlations between ToM and executive tasks ($F[13,168] = 1.2$, $p = .32$, $R^2 = 8.1\%$). This lack of evidence may be due to our limited focus on pairwise correlations, which cannot capture the joint contribution of different executive functions onto ToM competences. To account for this and similar effects, we thus performed the following series of multivariate linear regression analyses. In each age subgroup, we regressed performance scores in each ToM task onto performance scores in all executive tasks (and their two-way interactions). We evaluated both the in-sample fit accuracy and the cross-validation prediction accuracy (based upon a standard leave-one-out scheme). More precisely, we report both the standard R^2 and the "PRESS-R^2", which is based upon the predicted residual error sum of squares (PRESS) statistic (Tarpey, 2000). These analyses are summarized on Fig. 6B. First, note that the resulting sample size is one order of magnitude smaller than our previous analyses, since we had to select the participants that completed all executive tasks and at least one of the ToM

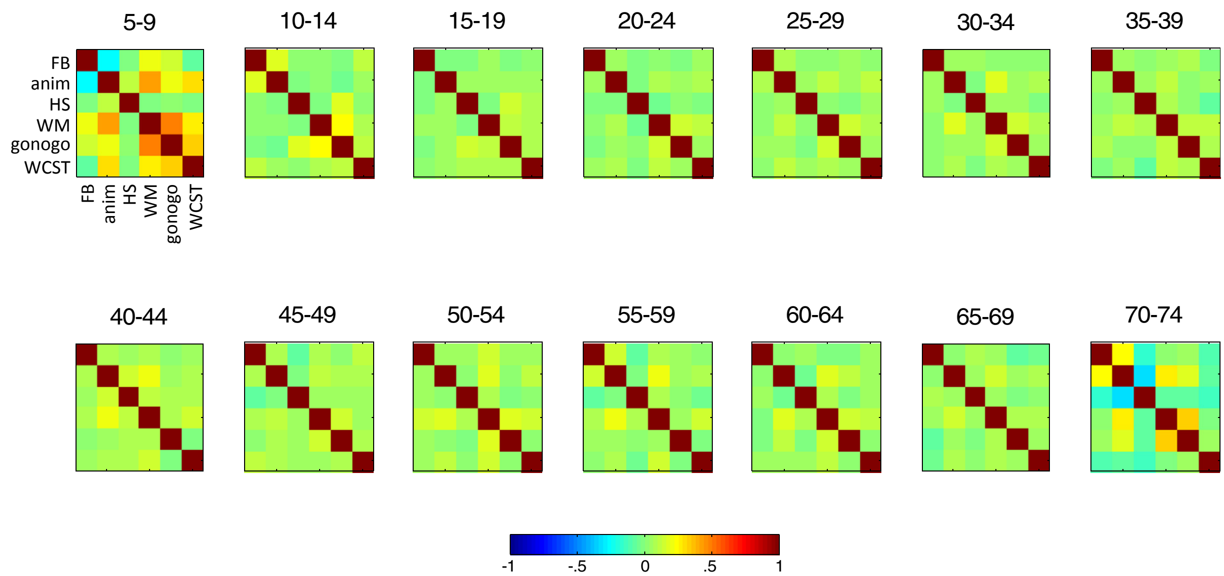

Fig. 5 – Non-parametric pairwise correlation analyses. Kendall's pairwise correlation matrices between performance scores (across participants) are shown for each age group. The color code ranges from −1 (blue) to 1 (red). In each panel, games are ordered as follows (from top to bottom and from left to right): *FB*, *anim*, *HS*, *WM*, *gonogo*, *WCST*. Note that we did not keep the two last age bins (above 75 years-old), because the corresponding correlation estimates were unreliable (some pairs of games were played by less than ten participants).

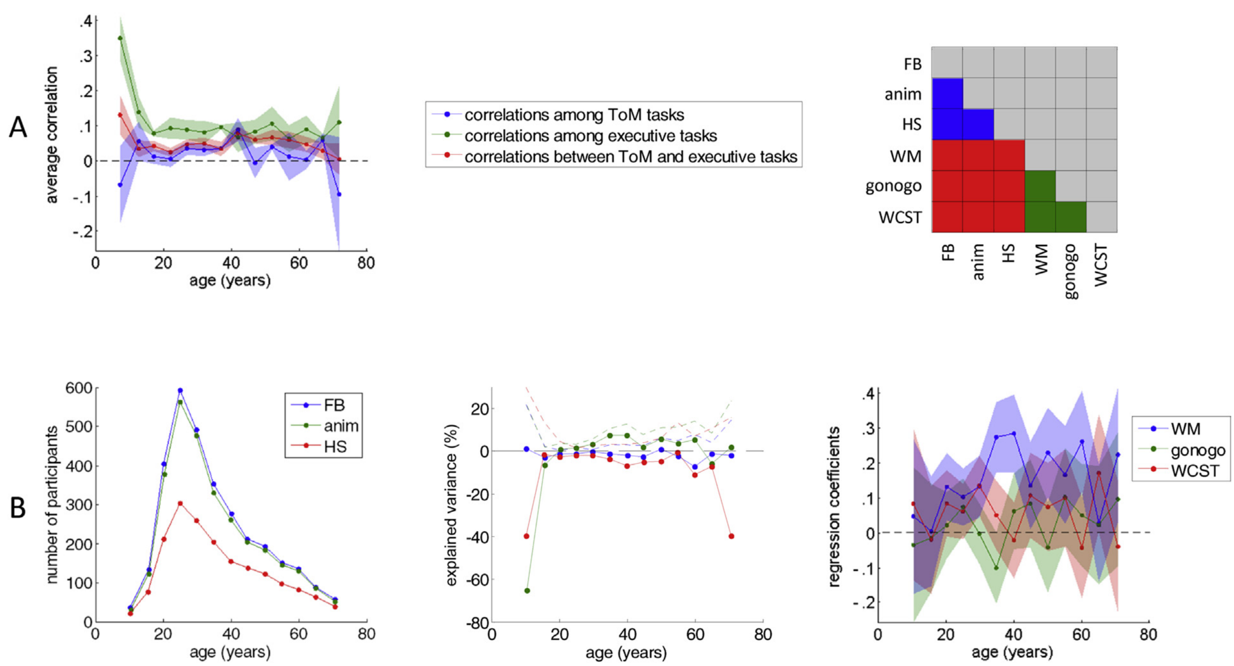

Fig. 6 – Temporal stability of cognitive interdependencies. A: Left: Average correlations (y-axis) are plotted as a function of age (x-axis), for each subset of Kendall's pairwise correlation matrices (blue: correlations among ToM tasks, green: among executive tasks, red: between executive and ToM tasks). Colored patches depict the 95% confidence intervals around the mean. Right: The associated partition of Kendall's pairwise correlation matrix is shown (identical color code). B: Multivariate regression analysis. Left: Number of participants who played all the executive games as well as one ToM game (y-axis) is plotted as a function of age (x-axis), for each ToM game (blue: *FB*, green: *anim*, red: *HS*). Middle: Percentage of explained variance (y-axis) is plotted as a function of age (x-axis), for each ToM game (same color code). Dashed lines indicate R^2 (model fit), whereas plain lines indicate PRESS-R^2 (cross-validation). Note that, in contrast to R^2, PRESS-R^2 can become negative (this happens whenever the magnitude of squared prediction error exceeds the sample variance). Right: Focus on anim. The estimated regression coefficients (y-axis) of each executive game (blue: *WM*, green: *gonogo*, red: *WCST*) are plotted as a function of age (x-axis). Colored patches depict the corresponding 95% Bayesian credible intervals.

games (*FB*: $n = 3137$; *anim*: $n = 2970$; *HS*: $n = 1782$). However, our participants' subsample is still representative of the population (when looking at the age profile on Fig. 6B, as well as the average performance scores – not shown –). The in-sample fit accuracy reproduces the qualitative U-shaped pattern of pairwise correlations seen on Fig. 5. But one can see that the high R^2 at early childhood and late adulthood are in fact overfitting artefacts, since the corresponding PRESS-R^2 falls below zero in all ToM games. In fact, PRESS-R^2 is always null or negative (!), except for *anim* at mid-adulthood. More precisely, in *anim*, PRESS-R^2 becomes positive at about 20 year old, peaks at about 40 years old (PRESS-$R^2 = 7.6\%$), and declines slowly until late adulthood. This means that only performance in *anim* can be reliably predicted from individual performances in executive tasks, and only between 20 and 60 years of age. We then asked what the relative contribution of each executive skill to performance in *anim* is. To avoid the overfitting issue we see in childhood and late adulthood, we fitted again the multivariate regression model using a Bayesian approach, with a tight prior variance on regression coefficients (set to the observed lifespan-variance of maximum-likelihood estimates). This essentially constrains the possible range of variation of fitted regression coefficients, eventually yielding lower out-of-sample generalization error (Liu & Aitkin, 2008; Scheibehenne & Pachur, 2014). The resulting 95% posterior credible intervals can be eyeballed on Fig. 6B. One can see that the only posterior credible intervals that exclude zero are those of *WM*, from about 20 to 60 years-old, which corresponds to the age range where PRESS-R^2 is positive. We will comment on the consistency of these results in the discussion section below.

4. Discussion

In summary, we have characterized the lifespan dynamics of cognitive development and decline for mentalizing competences, as well as for executive functions (i.e., working memory, inhibitory control and task switching). We have used behavioral tasks that engage mentalizing skills with increasing sophistication, from discriminating between one's belief and others' belief (*FB*), to recognizing others' intentions and emotions (*anim*), to predicting others' behavior in the context of strategic social interactions (*HS*). We have shown that although the maturation of executive functions occurs in synchrony (around 20 years of age), this is not the case for different ToM competences, which either mature before (low-level ToM) or after (high-level ToM) executive functions. This confirms that mentalizing may be decomposed into distinct underlying neurocognitive components. In line with this idea, we also have shown that distinct cognitive skills exhibit different relative decline rates, such that low-level ToM declines at a slower pace than executive functions, whose decline is slower than high-level ToM. In addition, we have shown that the statistical dependencies between ToM and executive skills change throughout the lifespan. In particular, we have shown that inter-individual differences in executive functions predict variability in ToM's skills from the onset of adulthood onwards (and only for higher-level ToM), i.e., after the complete maturation of executive functions.

Taken together, our results show a coherent picture, well aligned with the dual process theory of ToM (Frith & Frith, 2008). Recall that the synchronous maturation of standard executive functions justifies, from a developmental perspective, the notion of cognitive control as a homogenous neurocognitive system (Miller & Cohen, 2001). In this context, our critical finding is the fact that although similar levels of mentalizing performance are achieved before and after the maturation of cognitive control, executive functions significantly contribute to high-level ToM only after its complete maturation. This indicates that the cognitive architecture that underlies higher-level mentalizing competences may change over the lifespan. For example, differences in mentalizing skills among children and adolescents may be primarily driven by variations in linguistic skills, the development of which may scaffold early ToM maturation (Apperly et al., 2009).

Another, related, aspect of our results is the clear dissociation between low-level and high-level mentalizing competences. First of all, low-level ToM (i.e., the ability to discriminate between one's own and others' beliefs) matures much earlier than either cognitive control or high-level ToM. Second, it shows a very slow relative decline rate, when compared to cognitive control or higher-level ToM. In turn, only in low-level ToM (*FB*) does performance at late adulthood (above 65 years-old) remain significantly higher than at early childhood (up until 8 years-old). This dissociation would make sense if one would think of *FB* as a proxy for the representational component of ToM's dual process theory. In contrast, more complex games such as *anim* and *HS* would make heavy demands on cognitive control to reliably infer others' covert mental states from their overt behavior. This relates to the neural dissociation observed between spontaneous/automatic mentalizing and retrospective reasoning about others, where the former involves the temporo-parietal junction whereas the latter also engages the medial prefrontal cortex (Spiers & Maguire, 2006). Interestingly, neuroanatomical markers of the functional maturation of these brain systems also exhibit a pre-pubertal increase followed by post-pubertal loss, hence following an inverted U-shaped pattern over the lifespan (Sowell et al., 2003). Although existing neuroimaging studies vary in their estimation of the apex ages, maturation of the parieto-temporal junction has been shown to consistently precede the medial prefrontal cortex (Giedd et al., 1999; Gogtay et al., 2004; Paus, 2005; Sowell et al., 2003). This is important, since this provides a neuroanatomical footing to the dissociated dynamics of development and decline for low- and high-level mentalizing competences.

Nevertheless, one could argue that our dual-process interpretation does not explain the fact that inter-individual differences in cognitive control only predict variability in *anim*'s performances, and not in *HS*. This may be a side-effect of weak experimental control over the cognitive processes engaged in *HS*, which would also explain the weaker effect of age onto its performance (cf. Fig. 3). Under this view, task artefacts would dominate, and variations in *HS* performance would simply be due to chance. This is unlikely, however, given that *HS* and *anim* exhibit identical performance apices as well as relative decline rates (cf. Fig. 4). In fact, this reproduces the results of our previous investigation of mentalizing in *HS*

($n = 29$, age $= 22.5 \pm 3.8$ years), in which we showed that people's capability to outsmart artificial mentalizing opponents cannot be predicted from their performances in tasks assessing executive functions or empathy (Devaine, Hollard, & Daunizeau, 2014a). Rather, computational analyses of trial-by-trial choice sequences demonstrated that people's performance in *HS* critically depends upon whether or not they engage in recursive ToM inferences (of the sort "I believe that you believe that I believe…"). A possibility here is that such specifically social cognitive processes are not well captured by games such as *WM*, *gonogo* or *WCST*, hence compromising the statistical detection of the contribution of cognitive control to this form of high-level mentalizing.

Another related concern is the partial dissociation of *WCST* and other executive games in terms of their relative decline rates, despite the fact that all three performance apices are synchronous. Recall that the effect of aging onto cognitive decline is thought to be mediated by the loss of brain integrity (Allen, Bruss, Brown, & Damasio, 2005; Raz, 2000; Walhovd et al., 2005). This parallels the notion of "cognitive reserve" (Tucker-Drob, Johnson, & Jones, 2009; Zihl, Fink, Pargent, Ziegler, & Bühner, 2014), which relates to the brain's resilience to brain damage due to, e.g., compensatory processes. For example, it is known that the proportion of preserved functional network determines the extent of cognitive impairments in brain-lesioned patients (Nudo, 2013). Under this view, the rate of performance decline in any given task would decrease with the neuroanatomical specificity of its underlying neural bases (Greenwood, 2007). This would elucidate the quicker relative decline rate of *WCST*, which is known to engage broad frontal and parietal territories (Chan, Shum, Toulopoulou, & Chen, 2008; Nyhus & Barceló, 2009). Note that this would also explain why high-level mentalizing competences decline at a similarly quick pace, when compared to executive (prefrontal) tasks such as *WM* and *gonogo*. Extrapolating this neurocognitive scenario, one would expect that the increasing loss of brain integrity that inevitably occurs with aging eventually compromises the cognitive reserve, therefore limiting most forms of inter-individual variability. Interestingly, we indeed observed that inter-individual variability in performance seems to decrease (if any) with aging after about age fifty (not shown). Let us now discuss the main methodological aspects of our work.

First, we estimated age-related cognitive changes using a cross-sectional (as opposed to longitudinal) approach, which may be confounded by cohort effects. For example, one may argue that older adults have less experience with internet and/or smartphones, or that idiosyncratic attentional and/or motivational factors may confound age-related performance changes. We acknowledge that these and similar concerns may act as uncontrolled sources of variability in our data. In fact, both the absolute number of games played (a proxy for motivation) and the education level (in this context, a proxy for experience with internet) follow an inverted U-shaped pattern with age (data not shown). Nevertheless, recent studies have consistently demonstrated that cross-sectional web-based experiments and classical longitudinal studies yield similar estimates of cognitive development and decline (Dance, 2015; Germine et al., 2012; Hartshorne & Germine, 2015). Moreover, such cohort effects cannot explain the dissociations we see between low-level ToM, cognitive control and high-level ToM, both in terms of performance apices and relative decline rates.

Second, one may challenge our results on statistical grounds. For example, one may argue that, in our context, classical statistical significance testing may be inappropriate. This is because minuscule (i.e., behaviorally irrelevant) differences may eventually be deemed significant as sample size increases. One may then find unfortunate that, beyond mere statistical significance, the reported effects sizes are weak to moderate. However, they are in fact reminiscent of the effect of age onto neuroanatomical markers of development and decline, which barely reach thirty percent, even in much smaller samples (Allen et al., 2005; Walhovd et al., 2005). Note that the effect of age seems stronger for executive skills ($8.7\% < R^2 < 19\%$) than for mentalizing skills ($1.7\% < R^2 < 6.5\%$). We also did not justify the use of parametric statistics (e.g., ANOVA) when analyzing ordinal and/or binary performance scores. For completeness, we ran a similar logistic regression on *FB* data, which yielded qualitatively identical results (not shown). This was in fact expected under the central limit theorem, which guarantees that the average of a sufficiently large number of iterates of independent random variables will be approximately normally distributed, regardless of the underlying distribution. In fact, it is a well known fact in parametric statistics that inference in the context of very large datasets is very robust to the violation of normality assumptions (Lumley, Diehr, Emerson, & Chen, 2002). Similarly, differences in the number of subjects represented in each age bin may induce statistical imbalance in our between-subject design. But here again, ANOVAs and related parametric analyses are long-known to be immune to such form of imbalance (Draper, 2009; Shaw & Mitchell-Olds, 1993). One may also be puzzled by the apparent contradiction between the simple pairwise correlation analyses (cf. Fig. 5) and the multivariate regression analyses (cf. Fig. 6). This type of contradiction, however, often arises in the presence of multiple correlations between explanatory variables (Lawrance, 1976). In fact, none of the apparently high pairwise correlations were found to be significant, when correcting for multiple comparisons (not shown). Ultimately, cross-validation analyses, which are designed to avoid overfitting issues, demonstrated that these correlations were in fact due to chance. Regarding the heterogeneity of our statistical methods (classical significance testing, profile likelihood estimation, Monte-Carlo sampling, cross-validation), we would like to highlight that we aimed at striking a reasonable balance between simplicity, efficiency and robustness. Note that our analysis code is derived from the academic VBA freeware (Daunizeau et al., 2014): it is self-contained and can be made available upon request (please refer to the BRAiN'US website).

Let us now come back to the apparent poor test–retest reliability of *FB*. This is in fact highly surprising, given that this test has been shown to be highly reliable across a wide range of experimental procedures and age ranges (Ahmadi, Jalaie, & Ashayeri, 2015; Hughes et al., 2000). Recall that we included *FB* in our reliability analysis for completeness. However, we do not think that *FB*'s reliability can be challenged by poor test–retest repeatability in our context. In fact, a closer inspection of the test–retest data shows that evidence against repeatability is mostly driven by a subset of participants who first succeeded, and then failed (cf. Fig. 2). Most of these participants were adults above age 20. We admit that, given the nature of the test, we do not know why people who succeeded would want play the game twice. Nevertheless, we would argue that such participants were in fact exploring the alternative outcomes of the games, without caring for giving the right answer.

In conclusion, we have provided evidence for dissociation between low-level and higher-level mentalizing skills, both in terms of their dynamics of development and decline, and in terms of their dependence to cognitive control. These results provide a developmental validation of some key predictions of ToM's dual process theory (Frith & Frith, 2008).

Acknowledgements

JD acknowledges support from the IHU-a-ICM (Institute for Translational Neurosciences, Paris, France). The data reported in this study will be made entirely available on demand (please visit the BRAiN'US website). The authors are grateful to Pr. Uta Frith (and to Pr. Francesca Happé) for kindly supporting us when developing the app version of the game "Triangles at the Box Office".

REFERENCES

Ahmadi, S. Z. Z., Jalaie, S., & Ashayeri, H. (2015). Validity and reliability of published comprehensive theory of mind tests for normal preschool children: A systematic review. *Iranian Journal of Psychiatry, 10*, 214–224.

Aitkin, M. (2014). Profile likelihood. In *Wiley StatsRef: Statistics reference online*. John Wiley & Sons, Ltd.

Allen, J. S., Bruss, J., Brown, C. K., & Damasio, H. (2005). Normal neuroanatomical variation due to age: The major lobes and a parcellation of the temporal region. *Neurobiology of Aging, 26*, 1245–1260.

Apperly, I. A., Samson, D., & Humphreys, G. W. (2009). Studies of adults can inform accounts of theory of mind development. *Developmental Psychology, 45*, 190–201.

Aron, A. R. (2007). The neural basis of inhibition in cognitive control. *Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry, 13*, 214–228.

Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *NeuroImage, 5*, 49–62.

Brown, H. R., Zeidman, P., Smittenaar, P., Adams, R. A., McNab, F., Rutledge, R. B., et al. (2014). Crowdsourcing for cognitive science – the utility of smartphones. *Plos One, 9*, e100662.

Bull, R., Phillips, L. H., & Conway, C. A. (2008). The role of control functions in mentalizing: Dual-task studies of theory of mind and executive function. *Cognition, 107*, 663–672.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences, 12*, 187–192.

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

Chan, R. C. K., Shum, D., Toulopoulou, T., & Chen, E. Y. H. (2008). Assessment of executive functions: Review of instruments and identification of critical issues. *Archives of Clinical Neuropsychology, 23*, 201–216.

Cullen, H., Kanai, R., Bahrami, B., & Rees, G. (2014). Individual differences in anthropomorphic attributions and human brain structure. *Social Cognitive and Affective Neuroscience, 9*, 1276–1280.

Dance, A. (2015). News feature: How online studies are transforming psychology research. *Proceedings of the National Academy of Sciences, 112*, 14399–14401.

Daunizeau, J., Adam, V., & Rigoux, L. (2014). VBA: A probabilistic treatment of nonlinear models for neurobiological and behavioural data. *Plos Computational Biology, 10*, e1003441.

Devaine, M., Hollard, G., & Daunizeau, J. (2014a). The social Bayesian brain: Does mentalizing make a difference when we learn? *Plos Computational Biology, 10*, e1003992.

Devaine, M., Hollard, G., & Daunizeau, J. (2014b). Theory of mind: Did evolution fool us? *Plos One, 9*, e87619.

Draper, N. R. (2009). Analysis of messy data volume 1: Designed experiments. In G. A. Milliken, & D. E. Johnson (Eds.), *International statistical review* (2nd ed., Vol. 77, pp. 321–322).

Ferguson, F. J., & Austin, E. J. (2010). Associations of trait and ability emotional intelligence with performance on theory of mind tasks in an adult sample. *Personality and Individual Differences, 49*, 414–418.

Flobbe, L., Verbrugge, R., Hendriks, P., & Krämer, I. (2008). Children's application of theory of mind in reasoning and language. *Journal of Logic, Language and Information, 17*, 417–442.

Frith, C. D., & Frith, U. (2008). Implicit and explicit processes in social cognition. *Neuron, 60*, 503–510.

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology, 63*, 287–313.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of "theory of mind." *Trends in Cognitive Sciences, 7*, 77–83.

German, T. P., & Hehman, J. A. (2006). Representational and executive selection resources in "theory of mind": Evidence from compromised belief-desire reasoning in old age. *Cognition, 101*, 129–152.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review, 19*, 847–857.

Giedd, J. N., Blumenthal, J., Jeffries, N. O., Castellanos, F. X., Liu, H., Zijdenbos, A., et al. (1999). Brain development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience, 2*, 861–863.

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., et al. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences of the United States of America, 101*, 8174–8179.

Gordon, A. C. L., & Olson, D. R. (1998). The relation between acquisition of a theory of mind and the capacity to hold in mind. *Journal of Experimental Child Psychology, 68*, 70–83.

Greenwood, P. M. (2007). Functional plasticity in cognitive aging: Review and hypothesis. *Neuropsychology, 21*, 657–673.

Hartshorne, J. K., & Germine, L. T. (2015). When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychological Science, 26*, 433–443.

Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science, 344*, 1243091–1243091.

Hooker, C. I., Bruce, L., Lincoln, S. H., Fisher, M., & Vinogradov, S. (2011). Theory of mind skills are related to gray matter volume in the ventromedial prefrontal cortex in schizophrenia. *Biological Psychiatry, 70*, 1169–1178.

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.

Lawrance, A. J. (1976). On conditional and partial correlation. *The American Statistician, 30*, 146–149.

Lebreton, M., Kawa, S., d'Arc, B. F., Daunizeau, J., & Pessiglione, M. (2012). Your goal is mine: Unraveling mimetic desires in the human brain. *The Journal of Neuroscience, 32*, 7146–7157.

Lewis, P. A., Rezaie, R., Brown, R., Roberts, N., & Dunbar, R. I. M. (2011). Ventromedial prefrontal volume predicts understanding of others and social network size. *NeuroImage, 57*, 1624–1629.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology, 46*, 551–556.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology, 52*, 362–375.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health, 23*, 151–169.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*, 167–202.

Nelson, H. E. (1976). A modified card sorting test sensitive to frontal lobe defects. *Cortex Journal devoted to the Study of the Nervous System and Behavior, 12*, 313–324.

Nudo, R. J. (2013). Recovery after brain injury: Mechanisms and principles. *Frontiers in Human Neurosciences, 7*, 887.

Nyhus, E., & Barceló, F. (2009). The Wisconsin card sorting test and the cognitive assessment of prefrontal executive functions: A critical update. *Brain and Cognition, 71*, 437–451.

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

Paus, T. (2005). Mapping brain maturation and cognitive development during adolescence. *Trends in Cognitive Sciences, 9*, 60–68.

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a "theory of mind". *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*, 731–744.

Qureshi, A. W., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective selection, not level-1 visual perspective calculation: Evidence from a dual-task study of adults. *Cognition, 117*, 230–236.

Raz, N. (2000). Aging of the brain and its impact on cognitive performance: Integration of structural and functional findings. In F. I. M. Craik, & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (2nd ed., pp. 1–90). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Scheibehenne, B., & Pachur, T. (2014). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bulletin & Review, 22*, 391–407.

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science, 325*, 883–885.

Shaw, R. G., & Mitchell-Olds, T. (1993). Anova for unbalanced data: An overview. *Ecology, 74*, 1638–1645.

Sowell, E. R., Peterson, B. S., Thompson, P. M., Welcome, S. E., Henkenius, A. L., & Toga, A. W. (2003). Mapping cortical change across the human life span. *Nature Neuroscience, 6*, 309–315.

Spiers, H. J., & Maguire, E. A. (2006). Spontaneous mentalizing during an interactive real world task: An fMRI study. *Neuropsychologia, 44*, 1674–1682.

Tager-Flusberg, H., & Sullivan, K. (2000). A componential view of theory of mind: Evidence from Williams syndrome. *Cognition, 76*, 59–90.

Tarpey, T. (2000). A note on the prediction sum of squares statistic for restricted least squares. *The American Statistician, 54*, 116–118.

Tucker-Drob, E. M., Johnson, K. E., & Jones, R. N. (2009). The cognitive reserve hypothesis: A longitudinal examination of age-associated declines in reasoning and processing speed. *Developmental Psychology, 45*, 431–446.

Van Overwalle, F., & Vandekerckhove, M. (2013). Implicit and explicit social mentalizing: Dual processes driven by a shared neural network. *Frontiers in Human Neurosciences, 7*.

Walhovd, K. B., Fjell, A. M., Reinvang, I., Lundervold, A., Dale, A. M., Eilertsen, D. E., et al. (2005). Effects of age on volumes of cortex, white matter and subcortical structures. *Neurobiology of Aging, 26*, 1261–1270.

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.

White, S. J., Coniston, D., Rogers, R., & Frith, U. (2011). Developing the Frith-happé animations: A quick and objective test of theory of mind for adults with autism. *Autism Research: Official Journal of the International Society for Autism Research International, 4*, 149–154.

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.

Zihl, J., Fink, T., Pargent, F., Ziegler, M., & Bühner, M. (2014). Cognitive reserve in young and old healthy subjects: Differences and similarities in a testing-the-limits paradigm with DSST. *Plos One, 9*, e84590.