PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Review

Cite this article: Völter CJ, Tinklenberg B, Call J, Seed AM. 2018 Comparative psychometrics: establishing what differs is central to understanding what evolves. *Phil. Trans. R. Soc. B* **373**: 20170283.
http://dx.doi.org/10.1098/rstb.2017.0283

Accepted: 15 May 2018

One contribution of 15 to a theme issue 'Causes and consequences of individual differences in cognitive abilities'.

Subject Areas:
cognition, behaviour, evolution

Keywords:
individual differences, construct validity, executive functions, inhibitory control, comparative cognition, multi-trait multi-method test batteries

Author for correspondence:
Christoph J. Völter
e-mail: cjv3@st-andrews.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4154450.

THE ROYAL SOCIETY PUBLISHING

Comparative psychometrics: establishing what differs is central to understanding what evolves

Christoph J. Völter[1], Brandon Tinklenberg[2], Josep Call[1] and Amanda M. Seed[1]

[1]School of Psychology and Neuroscience, University of St Andrews, Westburn Lane, St Andrews, Fife, UK
[2]Department of Philosophy, York University, Toronto, Ontario, Canada

CJV, 0000-0002-8368-7201; BT, 0000-0003-3121-7328; JC, 0000-0002-8597-8336; AMS, 0000-0002-3867-3003

Cognitive abilities cannot be measured directly. What we can measure is individual variation in task performance. In this paper, we first make the case for why we should be interested in mapping individual differences in task performance onto particular cognitive abilities: we suggest that it is crucial for examining the causes and consequences of variation both within and between species. As a case study, we examine whether multiple measures of inhibitory control for non-human animals do indeed produce correlated task performance; however, no clear pattern emerges that would support the notion of a common cognitive ability underpinning individual differences in performance. We advocate a psychometric approach involving a three-step programme to make theoretical and empirical progress: first, we need tasks that reveal signature limits in performance. Second, we need to assess the reliability of individual differences in task performance. Third, multi-trait multi-method test batteries will be instrumental in validating cognitive abilities. Together, these steps will help us to establish what varies between individuals that could impact their fitness and ultimately shape the course of the evolution of animal minds. Finally, we propose executive functions, including working memory, inhibitory control and attentional shifting, as a sensible starting point for this endeavour.

This article is part of the theme issue 'Causes and consequences of individual differences in cognitive abilities'.

1. Introduction

The combination of the comparative method (i.e. comparing relevant traits across species) and analysing individual differences has proved to be a powerful approach to elucidate the evolution of physical traits. In principle, understanding the evolution of cognition can benefit from a similar approach. However, the study of cognitive evolution is complicated by the fact that cognition cannot be directly measured; instead, it must be inferred from measuring the physical substrate that underpins it (the brain) and its expression (behaviour) [1]. Initial theories of cognitive evolution were based on differences in relative brain size between taxa and correlations with various socio-ecological factors such as group size [2–4] and dietary diversity [3,5]. Although brain size is a crude index of cognitive ability [6], it does correlate with observational records of behaviour, such as tool use, social learning and innovation, in both mammals and birds [7,8]. Nevertheless, it has proved difficult to discriminate between theories for cognitive evolution based on this indirect evidence [9]. A satisfying account of cognitive evolution must describe the trait that is evolving more precisely.

Comparative psychologists have attempted to measure species differences in cognition more directly, by conducting experiments on different species either in the field or in the laboratory (for recent reviews, see [10,11]). Though this brings us one step closer to measuring cognition, experimental psychologists have long recognized that species differences in performance

on cognitive tests could result from multiple sources [12]. Such contributing causes include species differences in perception, temperament, motor control, body morphology and domain-general cognitive abilities that are peripheral to the targeted cognitive ability [13]. Additionally, cohort differences in experience or demographics can complicate species comparisons, particularly on a single task. Control conditions (administered within the same species and cohort) can help us ascertain that the results are not solely ascribable to these peripheral factors, though it is hard to be exhaustive in ruling out alternative causes for species differences in this way. Another approach is to seek positive evidence that the test is measuring the targeted ability through examination of individual differences. In other words to explore whether or not the cognitive ability can be shown to contribute to performance across different tasks.

In recent years, there has therefore been a shift towards examining individual variation in performance across multiple tasks rather than group performance in a single task [14–16] (though the importance of individual variation has been recognized for a long time, e.g. [17]). To that end, researchers have designed and administered test batteries to identify factors underlying individual differences in task performance. Most of these studies to date have been concerned with the question whether a common factor, commonly referred to as *g*, can be identified that accounts for between-subject variance across different tasks. Many comparative studies have found evidence for such a factor; others have not (for a recent review, see [18]). There are many possible explanations for the existence of *g* and we agree with others who have argued that evidence for a psychometric *g* factor does not entail the presence or absence of an overarching, domain-general reasoning ability that can be deployed for very diverse purposes [19–21].

In this paper, one of our goals is to analyse different approaches to test battery design with a view towards promoting a more systematic approach. We argue that it is time to go beyond the question whether or not *g* can be identified and advocate a three-step programme for designing test batteries that can elucidate the structure of cognition (i.e. what are the dissociable components of cognition and in what way are they related with one another?). To this end, the question of measurement is of central importance. Which cognitive abilities can be validated across different contexts (e.g. different behavioural tasks) and measured reliably across time? Little work to date has been dedicated to this important question. Giving more attention to validating cognitive abilities (see box 1 for the definition and discussion of test validity) will allow us to answer more detailed questions regarding the causes and consequences of individual differences in cognition [16,29]. Rather than looking for socio-ecological correlates of *g*, one can look for correlations between particular cognitive abilities and certain socio-ecological variables [12]. An example for such a targeted approach is the correlation between inhibitory control measures and fission–fusion dynamics across different primate species [23]. This kind of process-oriented approach could also be used within species to study the consequences of individual differences. While some recent evidence suggests that problem-solving abilities are related to fitness [30–32], other studies have not found this association [33]. However, it is largely unclear which (or indeed whether [13]) cognitive abilities underlie successful problem-solving in these cases. Uncovering whether or not variation in individual performance reliably measures variation in a certain cognitive ability is, we argue, a logical precursor to interpreting correlations with fitness (or lack thereof). If individual differences in task performance result largely from differences in experience or motivation, such differences are unlikely to be related to fitness, as they are transient. Conversely, if individual differences track body condition or health, correlations with fitness might be expected because of some interaction with this third variable, whether or not the task is a valid measure of a certain cognitive ability. Finally, with information about which cognitive abilities can be identified and how they are related to one another (i.e. the structure of cognition), one could start to ask further questions about how cognition evolves: for example, are certain abilities likely to undergo correlated evolution, or might they be traded off against one another?

2. Targeted test batteries: from *g* to more specific questions

The first step in test battery design is to specify the cognitive abilities to be tested. The second step is to specify what tasks are supposed to measure these abilities. Both trait (e.g. short-term memory) and task (e.g. finding food under a cup after a delay) selection will probably affect the latent variable structure supported by the study. Despite its influence on the generalizability of the results, often little justification of the trait and task selection is provided, especially in test batteries looking for a *g* factor. An unbalanced task selection, for example with a bias on learning tasks or spatial cognition tasks, might limit the conclusions that can be drawn from the results [16,29].

Two main approaches guiding trait and task selection can be identified. The first, which we have labelled the 'ethological' approach, is based on a careful analysis of a species' socio-ecological challenges and its typical behavioural solutions to these challenges. We will review an example for this approach, the primate cognition test battery (PCTB) [34]. The benefit of this approach is that it provides researchers with a good starting point for the design of ethologically valid tasks that are likely to tap into survival-relevant cognitive abilities. However, these tasks have usually not been designed with the explicit goal of investigating correlations of between-subject variation in performance. This can result in tasks that do not yield large between-subject variation, for instance, due to ceiling or floor effects or due to an insufficient number of trials per individual. The second, 'psychological' approach to test battery construction, is based on previous cognitive studies with the same or different species (e.g. based on the human psychometric literature). This approach is anchored in specific hypotheses about cognitive abilities which guide task selection criteria. For example, these hypotheses might specify response profiles (i.e. how an individual's performance is affected by different experimental manipulations) or error patterns in performance that the candidate tasks should provoke. In practice, the two approaches can overlap. Irrespective of the approach taken, using an established test battery for a different species might require significant task adjustments. Initial experimentation is essential to ensure that the tasks are suitable for the species of

Box 1. Establishing test validity.

Content validity

The starting point for establishing content validity is to determine the nature of a cognitive ability on theoretical and/or empirical grounds. In other words, researchers need to agree on features defining the ability of interest [22]. Often the rich body of cognitive research with humans can provide initial guidelines, especially in cases with limited pre-existing comparative research. The aim is to make predictions about response profiles, error patterns or signature limits that are specific to the cognitive ability under investigation. Based on these considerations, researchers can design experiments in which they manipulate task complexity to reveal the hypothesized response profiles and signature limits within individuals. At the group level, these signature limits should also be evident in comparison with control conditions that do not tax the targeted cognitive ability to the same extent as the test condition.

In the realm of executive functions (EFs), such content validity criteria include susceptibility to task interference in working memory tasks and switch costs in attentional set-shifting paradigms. Inhibitory control measures should yield response profiles and signature limits indicative of a prepotent response. Variability in inhibitory control can only be measured when there is some prepotency or interference that needs to be overridden. Prepotency, however, can only be shown when individuals at least occasionally make mistakes (e.g. when individuals of a species occasionally bump into a transparent cylinder when they try to reach a reward inside the cylinder). Ceiling (or floor) effects in performance make it impossible to establish content validity (e.g. great ape species performed close to or at ceiling in inhibitory control tasks including the cylinder and the A-not-B error task [23–25]). Ideally, researchers can design experiments that manipulate the task complexity with respect to the signature limits they are interested in. For inhibitory control tasks, this can be realized by manipulating the strength of the prepotent response. For example, in a go/no-go paradigm, increasing the relative frequency of go trials should negatively impact on the no-go performance [26].

Another strategy to reduce interpretational ambiguity is to focus on error patterns. Often mistakes can be more informative than success [27]. However, even if the performance is not at ceiling, errors might be related to factors other than the prepotent response (e.g. motivation and distractibility). Sometimes the task design can mitigate this problem by including various opportunities for making mistakes. The type of mistakes may hint towards different underlying causes. An example of such a task design is the A-not-B error task with three aligned cups: repeated exposure to hiding events of a target object under cup A can induce a search response towards cup A even when in probe trials the target object is hidden under cup B in full view of the participant (the so-called A-not-B error). Adding a third cup (C) to the set-up that is never used as a hiding place allows for distinguishing between inhibitory control errors (cup A) induced by the previous exposure to cup A hiding events and unspecific mistakes (cup C) in probe trials (when the target object is in cup B). In this way, adding different options for making mistakes will improve the task design by allowing the content validity of the task to be assessed.

Construct validity

Construct validity aims at triangulating variables (constructs) that account for variance in task performance [22]. Multiple tasks aiming at the same ability (but differing in peripheral task demands and stimulus appearance) should ideally produce shared variance in performance attributable to a common factor (convergent validity). Conversely, tasks that aim to measure different traits should not produce highly shared variance (discriminant validity). Convergent and discriminant validity together bolster construct validity. Based on shared variance across tasks alone, it remains unclear what the shared variance actually represents (e.g. general intelligence or a more specific cognitive ability). Multiple traits should therefore be examined within the same test battery to tease out what is shared and what is distinct, to discriminate and label latent variables. In this way, a multi-trait multi-method approach can help to establish construct validity and to elucidate the structure of cognitive abilities [28].

Correlating task performance at the species or group level is not sufficient for establishing construct validity at the individual level (within each species of interest). This is because species may differ in their performance in multiple tests due to differences in another variable (such as motivation), leading to correlated performance at the group level but not at the individual level.

interest, especially when a test battery is transferred to a distantly related taxon (e.g. from primates to birds or fish [35,36]).

In the next sections, we will briefly review the results of the 'ethological' and 'psychological' approaches to test battery construction with examples from studies of primate cognitive evolution. Following this analysis, we will propose some guiding principles for test battery design that arise from evaluating the strengths and weaknesses of the work to date.

(a) The primate cognition test battery

Herrmann *et al.* [34] initially designed the PCTB to compare the cognitive abilities of different great ape species (2.5-year-old human children, chimpanzees (*Pan troglodytes*) and orangutans (*Pongo pygmaeus*)) at the group level. In line with the 'ethological' approach, the design of the test battery was based on a review of the primate cognition literature examining challenges from the physical and social cognitive domains [37]. Although task design in the PCTB was anchored in the challenges faced by these species in their daily lives (to find and locate food, use tools and deal with conspecifics), the hypothesis being tested had to do with the structure of the underlying cognition: namely that social cognition would be dissociable from physical cognition and capable of evolving separately. The group-level analysis revealed some support for this notion, because 2.5-year-old

children outperformed chimpanzees and orangutans in the social cognitive domain but performed similarly to chimpanzees in the physical cognitive domain. However, as described above, these species differences in some tasks but not others could still, in principle, be the result of non-cognitive species differences [38,39].

Analysis of individual differences in performance can be used to further investigate the structure of cognition. In a later reanalysis of the original dataset of chimpanzees ($N = 106$) and children ($N = 105$) [40], between-subject variation in performance was examined even though the PCTB was not designed for this purpose. A confirmatory factor analysis (CFA) did not endorse the original division of the test battery into a social and a physical cognitive domain, though it did yield dissociable components. Instead, a spatial cognition factor could be identified in both chimpanzees and human children. For children, there were two additional factors, one that included the shared variance of some of the non-spatial physical cognition tasks (i.e. tool use and numerical cognition) and one capturing individual variation in social cognition tasks. For chimpanzees, there was one additional factor onto which some of the social and (non-spatial) physical cognition tasks loaded.

In a replication of the PCTB with another chimpanzee sample ($N = 99$), Hopkins *et al*. [41] found a different underlying structure using a principal component analysis (PCA): all spatial cognition tasks loaded on component 1, tool-related tasks and causal reasoning tasks loaded on component 2 and all social cognitive tasks loaded on a third component. Hopkins *et al.* also found evidence for test–retest reliability of the measures, though there was some improvement in the spatial and numerical cognition tasks over time. They also found evidence for a common *g* factor. Moreover, they found evidence for heritability of this composite score.

Multiple differences could account for inconsistent latent variable structures between the two chimpanzee studies: apart from differences in the statistical analyses (CFA versus PCA), the later study by Hopkins *et al.* [41] included only 13 out of the original 15 tasks. One of these tasks, the addition task, loaded on the physical–social factor identified by Herrmann *et al*. [40]. It is therefore possible that the inconsistent results might be due to methodological differences between the studies rather than differences between the two chimpanzee samples. This highlights the vital importance of task choice and inclusion in test battery design. Nevertheless, the results from the PCTB yielded patterns of correlation and dissociation that provided some evidence for construct validity (box 1), in particular for spatial cognition, although the details of cognitive mechanisms underlying these common factors remain opaque.

(b) Beyond *g* and social/physical cognition?

In the following, we will review some studies that have been anchored in a more 'psychological' approach to test battery design. Rather than starting from ethology and using factor analysis to examine the structure of the underlying cognition, these studies start from hypotheses about the nature of cognition and have compared species that inhabit different socio-ecological niches to examine hypotheses about the evolutionary causes/consequences of differences in specific cognitive abilities. Apart from the aforementioned studies of *g*, most of the studies using this approach have focused on inhibitory control.

(c) Inhibitory control

Inhibitory control is often described as a component of EFs, a suite of domain-general, partially independent cognitive abilities that are important in maintaining goals even in the presence of interference and switching flexibly between goals [42–44]. Inhibitory control (for a critique of this term, see [45]) includes response inhibition and interference control. Response inhibition refers to the top-down capacity to suppress a (stimulus-driven) prepotent response and/or to activate another (memory-based) response instead. Interference control refers to the ability to focus on goal-relevant information in the presence of distracting information (for a recent overview article on the terminology and definitions, see [44]).

Such a domain-general cognitive ability should lead to consistent individual differences across different contexts. In the human literature, the evidence for such a common cognitive ability is mixed [46–50]. A large-scale meta-analysis (based on 282 samples and over 33 000 participants) examined convergent validity of self-control measures with human adults [51]. They found modest convergence of measures of self-control (defined here as 'top-down processes that inhibit or obviate impulses' [51, p. 260]) with informant-report and self-report questionnaires yielding the highest convergence scores and executive function tasks (the most frequently used tasks were go/no-go, Stroop and set-shifting paradigms) exhibiting smaller, yet significant convergent validity with other EF tasks (average Pearson's correlation coefficient among EF tasks: $r = 0.15$). There was also a significant correlation among delay of gratification tasks (DoG; average correlation coefficient: $r = 0.21$) but notably no significant correlations between DoG and other EF tasks (average correlation coefficient: $r = 0.11$). It has been suggested that temporal discounting makes DoG tasks different from other inhibitory control tasks [10]. Temporal discounting refers to the degradation of the subjective value of a reward with increasing delays. Individual differences in temporal discounting might therefore be supported by different cognitive processes [44,52]. Indeed, investigations of the neural correlates of response inhibition and choice impulsivity (or temporal discounting) cast doubt on whether the same cognitive processes are involved here [53], which would explain the inconsistent individual differences between DoG and other self-control tasks.

Nevertheless, this meta-analysis is consistent with the notion of some common cognitive ability underpinning tasks aiming at measuring inhibitory control in humans, though it seems unlikely to be unitary. In fact, it has been proposed that inhibitory control can be further decomposed into three subcomponents including stimulus detection, action selection and action execution [54]. Whether this common factor should be labelled inhibitory control (or a suite of inhibitory control abilities) depends on whether discriminant validity (box 1) can be established with other constructs such as general intelligence, shifting and working memory. An assessment of individual differences in EF in humans showed that tasks aiming at measuring inhibitory control did not load onto an independent factor but on a common factor of EF [43]. It therefore seems questionable to treat all

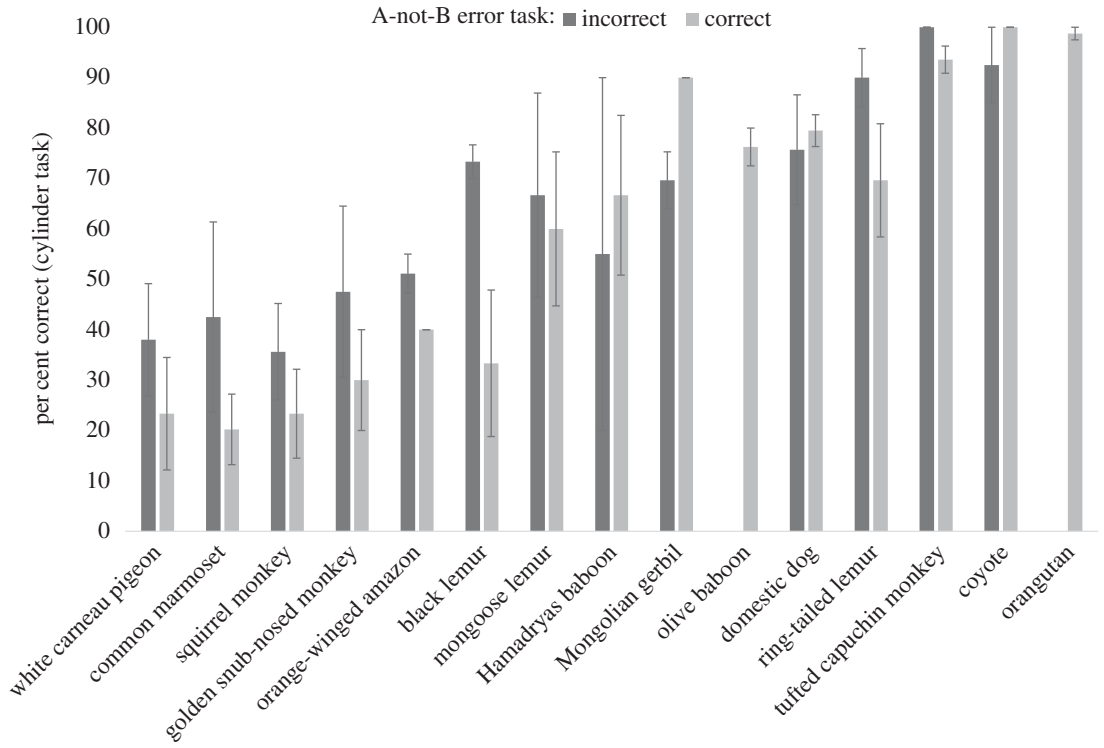

Figure 1. Mean performance (± s.e.) in the cylinder task (per cent correct out of 10 trials) of the MacLean *et al*. [24] dataset as a function of species (ranked by overall performance in the cylinder task) and performance in the A-not-B error task (correct/incorrect, 1 trial). Only the performance of individuals that completed both tasks is shown.

of the tasks entered in this meta-analysis (such as set-shifting paradigms) as primary measures of inhibitory control (we discuss the task impurity problem further below).

In the comparative literature, two fairly large-scale studies have administered cognitive test batteries that were explicitly designed to compare species in their inhibitory control ability [23,24]. However, without first establishing convergent and divergent validity of the deployed inhibition measures, it is unclear if this would assay one ability or several. Amici *et al*. [23] presented five tasks aiming at inhibitory control to six different primate species. Some of these tasks were based on classical psychological tasks of inhibitory control (e.g. the A-not-B error); others were somewhat more ethologically grounded, such as detour-reaching tasks. They found an association between performance on these tasks and sociality, with species that have a more fluid social structure (fission–fusion) performing better on average than those with less complex social organization. MacLean *et al*. [24] presented two inhibitory control tasks (A-not-B and the cylinder task, a detour-reaching test) to 567 individuals representing 36 species with a wide phylogenetic coverage. They found a correlation between test performance and absolute brain volume (but see [55,56] for recent evidence with corvids questioning this finding) and, for primates, an association with ecology (dietary diversity). Implicit in the rationale for both studies is that inhibitory control is a unitary, domain-general ability undergoing evolutionary change, though the results point to different selection pressures at work. However, although performance in the A-not-B error and cylinder task of the MacLean *et al*. dataset was correlated at the species level (when controlling for phylogeny [24]), individual differences were not examined. In the following, we examine these datasets from an individual differences perspective, to explore the evidence to date for construct validity of inhibitory control in the comparative framework (box 1). However, it should be noted that only three of the species in the Amici *et al*. dataset (spider monkeys, *Ateles geoffroyi*, $N = 18$; capuchin monkeys, *Sapajus apella*, $N = 27$; long-tailed macaques, *Macaca fascicularis*, $N = 12$) received more than two trials per task and can provide therefore measures of individual variance; similarly, only a subset of individuals in the MacLean *et al*. study received both tasks ($N = 216$).

To test the convergence of the two measures administered by MacLean *et al*. [24], we modelled the influence of accuracy in the A-not-B error task (1 trial: correct/incorrect) on the performance in the cylinder task (number of correct responses: 0–10). If the two tasks measure individual differences in the same cognitive ability, i.e. response inhibition, one might predict that individuals who choose the correct cup in the A-not-B error task would also perform well in the cylinder task (compared with individuals who committed the A-not-B error). We included all tested species from the MacLean *et al*. dataset with more than six individuals participating in both tasks (in total 192 individuals representing 15 species, figure 1). We used a Poisson generalized linear mixed model (GLMM) to analyse the number of correct trials in the cylinder task and added A-not-B error task performance as the predictor variable and species and subject ID as a random effect (see the electronic supplementary material for more information on model assumptions and the detailed model output). We found that the A-not-B error task was not significantly associated with the cylinder task performance (GLMM 01: $\chi^2(1) = 2.13$, $p = 0.145$). When we excluded two species that exhibited ceiling effects in the A-not-B error task (orangutans and olive baboons), the p-value was slightly smaller (GLMM 02: $\chi^2(1) = 3.28$, $p = 0.070$). However, if anything, individuals that successfully located the food in the A-not-B

error task tended to perform worse in the cylinder task (according to the model by 18%) compared with those that committed the A-not-B error (estimate $\pm$ s.e: -0.19 ± 0.11, 95% CI [-0.41, 0.02]). The overall pattern of results did not change when we only analysed the 10 primate species (GLMM 03). The power to detect an estimate of 0.3 (corresponding to a 35% increase in cylinder task performance) for the predictor A-not-B Error Task Performance was 81.2% for GLMM 01 (GLMM 02: 73.9%; GLMM 03: 61.2%; see the electronic supplementary material for details). In line with these results, a previous study in dogs (*Canis familiaris*, $N = 30$) focusing on individual differences did not find evidence for convergent validity of these two tasks [57]. Similarly, a recent study with pheasants (*Phasianus colchicus*, $N = 81$) found no evidence for correlated performance in two detour tasks involving transparent materials blocking direct access to a food reward [58]. Moreover, the content validity of the A-not-B paradigm has recently been challenged, as hand-tracking training but not experience with another inhibitory control task (reversal learning) substantially improved the A-not-B error task performance of New Caledonian crows (*Corvus moneduloides*) [56].

In the Amici *et al*. dataset [23], we correlated individual performance across four different inhibitory control measures (see the electronic supplementary material). The sample sizes for each pairwise comparison ranged from 5 to 19 depending on the task comparison and the species. The small sample sizes and resulting low power (7–24% assuming a medium effect of $r = 0.3$; see the electronic supplementary material) would therefore not allow us to detect weak to moderate correlations. To detect medium effects with a satisfactory power of 80%, a sample size of 84 individuals would be required. However, we expected to find at least positive correlation coefficients if these tasks measured the same underlying ability. We found no such simple picture. In particular, the delay of gratification (DoG) task did not seem to be related consistently with the other inhibitory control measures within any of the different species (five out of nine correlation coefficients were negative). Indeed, there was even a significant negative correlation between DoG and the middle-cup task for the spider monkeys. The other tasks showed more consistent individual differences within each species (eight out of nine correlation coefficients were positive across species, but only two of these correlations were also statistically significant; see the electronic supplementary material). These tasks relied on prepotent responses induced by tendencies to either reach for visible food directly (Plexiglas hole), repeat previously rewarded choices (A-not-B error) or follow a proximity bias when searching for hidden food rewards (middle cup). The lack of relationship between DoG and other inhibitory control measures in this study is consistent with the above-mentioned meta-analysis of the EF literature with human adults, which found no clear pattern of correlation between DoG tasks and other EF tasks (including classic response inhibition tasks [51]).

From our analysis of the studies on inhibitory control, we have found little evidence for convergent validity in this cognitive domain. Importantly, even if the multiple measures had yielded correlated performance, it would remain unclear what the shared variance represents. For example, even if we had found evidence for convergence of different inhibitory control measures, we still would not know whether we could attribute the shared variance to inhibitory control or another ability. The lack of convergent validity of different inhibitory control measures might be attributable to low statistical power (especially in the case of the Amici *et al*. dataset) or masking effects of confounding factors (other cognitive abilities, non-cognitive factors including motivation, etc.), or it might indicate that inhibitory control is not a unitary ability. Only the multi-trait multi-method approach with sufficient sample sizes could help to mitigate this interpretational challenge by establishing both convergent and discriminant validity. Following the human literature on EF, multiple measures per trait (e.g. updating, shifting and inhibition) would be needed [43]. As we have seen in this section, further task development for non-human animals is required for the assessment of inhibitory control. In the final section, we discuss how the multi-trait multi-method approach might be implemented in comparative psychology.

3. The structure of cognition: a problem with two unknowns

Our review of the literature has identified several pitfalls in the use of test batteries to elucidate the structure of individual differences from a comparative perspective, even when studies focused on a single trait (e.g. inhibitory control) measured by multiple tasks. One of the main challenges for the investigation of individual differences in cognitive abilities is that we start with two unknowns: first, we do not know what our tasks measure (the so-called task impurity problem; see [59]) and, second, we do not know which cognitive traits exist and can be measured within a species (the construct validity problem).

All behavioural tasks are 'impure', in the sense that it is not possible to isolate and measure a cognitive ability with a single task. Confounding factors that contribute to task performance are other cognitive abilities (apart from the target ability) and non-cognitive factors including motivation, personality traits [60,61] and prior experience [13]. To complicate matters even further, the cognitive and non-cognitive factors that contribute to task performance might vary between individuals, and they may not lead to stable effects across time. For example, over time some individuals, unlike others, might adopt strategies to cope with the task demands more efficiently which in turn will affect the cognitive load of the task for these individuals.

We advocate a three-step solution to tackle these problems:

1. First, establish *content validity*. Does performance on the task accord with theoretical principles underlying the hypothesized ability? There are two main tools to examining this: firstly, signature limits in performance, and secondly systematic variation across conditions. Signature limits refers to the way individuals make mistakes (including commission and omission mistakes). Analysing these error patterns can help to establish the content validity of a task [27]. Systematic variation refers to initial experimental work showing at the group level that the test condition differs from control conditions in meaningful ways (as predicted by the targeted ability). Importantly, tasks that reveal such signature limits or systematic variation need to be established for every species under study, as a task that has demonstrable content validity for one species might not be appropriate for another.

2. The second step is to assess the repeatability (or test–retest reliability) of candidate measures. Only tasks that yield a consistent ranking of individual performance over time are good candidates for capturing cognitive abilities. Depending on the ability under investigation, learning effects might hinder the assessment of test–retest reliability (e.g. due to ceiling effects in the retest). Changing the task-relevant stimuli between test and retest can help to remedy this problem.
3. Third, valid and reliable measures of individual variation in a cognitive ability can be combined in a multi-trait multi-method test battery to deal with task impurity and the construct validity problem [28]. The aim of such a latent variable approach is to establish *convergent* and *discriminant* validity, the constituents of construct validity (box 1).

Each of these steps aiming at content validity, repeatability and construct validity will require intense and possibly coordinated research effort. Our contention is that these steps should be undertaken in order for a study of individual differences in cognition to be maximally meaningful. Fortunately, each of the steps constitutes interesting research questions in its own right.

One complication concerns the optimal choice of tasks at each step. Step 1 is made easier with robust effects (replicable at the group level) that can support statistical tests between different conditions. However, there is no guarantee that a valid and robust test of an ability will yield individual variation in that ability across individuals, which is needed for Steps 2 and 3. In fact, the most robust tests may not translate into reliable measures of individual differences precisely because they tend to be associated with small between-subject variance [48]. Moreover, it is important to consider what kind of dependent variables are extracted from the task. Difference scores (e.g. test condition performance subtracted by control condition performance) can have a lower signal-to-noise ratio compared with their constituents (e.g. the test condition performance) and therefore might not provide sensitive measures [48,62]. Nevertheless, difference scores might remove systematic between-subject variation in performance unrelated to the cognitive ability under investigation (which is desirable for Step 3). Having the three-step programme in mind at the outset of task design is therefore beneficial for future-proofing tasks, for example by exploring multiple levels of difficulty in Step 1, to allow difficulty to be titrated at Steps 2 and 3 to avoid floor and ceiling effects.

4. Future directions for the psychological approach

We deem EFs a good starting point for the assessment of individual differences in behavioural flexibility for multiple reasons: EFs are thought to be domain-general processes that affect the performance in most behavioural tasks. Twin studies suggest that individual differences in EFs in humans are almost entirely of genetic origin [63]. Moreover, EFs are correlated with mental and physical health measures in humans (for a review, see [42]) and survival (in the context of chronic illness [64]). Thus, EFs might also be the ideal candidate for looking into causes and consequences of individual differences in cognition.

To date, most research attention has been devoted to three skills that together are thought to represent the pillars of EF: working memory updating, attentional shifting (also known as set-shifting or cognitive flexibility) and inhibition (including response inhibition and interference control). The multi-trait multi-method approach has been applied to study the structure of individual differences in EFs in humans, including multiple tasks aiming at updating, shifting and inhibition [43,59]. According to one of the most influential models of human executive functions by Miyake & Friedman [65], there is a common factor onto which all of these tasks load. Additionally, there are two nested factors, an updating-specific factor and a shifting-specific factor, that represent the shared variance unique to the updating and shifting tasks. The shared variance of the inhibition tasks, however, cannot be differentiated from the common EF factor. In human preschool children, in contrast, a single factor seems to be sufficient to account for individual differences across EF tasks [66,67]. The latent variable structure underpinning individual differences in cognition might thus be subject to developmental change. Systematic differences in the age structure of different study samples can therefore not be neglected when different species are compared.

One might argue that identifying the latent variable structure of performance on EF tasks does not eliminate reference to control homunculi or black boxes [54]. While this is true, identifying such latent variable structure might serve as an intermediary step towards a more mechanistic model of EF [68]. Breaking down EF into its fundamental components will probably require iterative applications of the multi-trait multi-method approach. Conversely, inspiration for task designs and task selection can also be drawn from existing computational or mechanistic models of EF (e.g. [54,69,70]). For example, models that decompose executive control of actions further (e.g. into signal detection, action selection and action execution [54]) can help to make predictions about response profiles and signature limits. Besides, such models might help to explain why convergent validity of inhibitory control measures has proved hard to establish. Any of the proposed action control subcomponents (or a certain combination thereof) might explain individual differences in task performance.

In the comparative literature, individual differences in EFs have not been systematically investigated [71]. There are some notable exceptions linking *g* to working memory performance in mice [72,73]. Moreover, a meta-analysis reported in this issue provides evidence for low to moderate convergent validity and test–retest reliability estimates for a number of different tasks (including inhibition and reversal learning tasks) and species [74]. In most taxa, however, research looking into the structure of individual differences in EF is missing. Fortunately, there are a number of paradigms that have been used to tax different EFs, including working memory (e.g. [75,76]) and inhibitory control tasks (e.g. [23,24,77,78]). The first steps towards a psychometric examination of EF in non-human animals will be to establish the content validity and reliability of these paradigms in different species.

Valid measures of EFs will also help us to interpret individual differences in more specific domains. Most behavioural tasks, especially the ones that require a change in

behaviour, arguably will at least initially tax EFs to a varying degree (but increasing experience with task-relevant contingencies can lead to automatized control processes due to learning [44,54,79]). It is therefore important to examine the extent to which observed differences in task performance are due to differences in EF. For instance, it has been suggested that the development of EF over the preschool years may constrain, or enable, the emergence of abilities such as theory of mind and object permanence [80,81]. Interestingly, it is possible that a similar argument could apply over a phylogenetic time scale [82].

5. Conclusion

Investigating the structure of individual differences in cognitive performance within a species will lead to insights into the causes and consequences of individual variation, and it will allow for more informative comparisons across species. To this end, we need to refine the assessment of individual differences in behavioural flexibility. Following a classic psychometrics approach, we advocated a three-step programme: first, experimental work to establish paradigms that yield response profiles indicative of the targeted ability; second, assessments of reliability of individual differences across time; third, multi-trait multi-method test batteries to establish validity of the targeted ability. Elucidating the structure of cognition across different species will be a challenging endeavour. One of the biggest obstacles will be to obtain sufficient sample sizes. For many species (including most primate species) that are difficult to access and whose sample sizes in captivity are usually small, the only remedy will be large-scale collaborative projects across laboratories or field sites. This will certainly be no easy feat especially because such expensive, long-term projects are difficult to realize in an academic environment with short-term funding, but there are encouraging examples in related research areas that show the feasibility of such projects (e.g. the ManyBabies project [83]). A first pilot project aiming at establishing large-scale collaboration in the field of comparative cognition is currently underway (the ManyPrimates project [84]).

In this article, we cautioned against taking short-cuts when constructing test batteries. Given the amount of work that is necessary to conduct a test battery with a sufficiently large number of individuals, a trial-and-error approach cannot be recommended. Borrowing the test battery design from research with another species will probably result in biased outcomes; in the worst scenario, it might lead to ceiling or floor effects. New task designs and pilot work to establish certain response signatures (the content validity) within each species are advisable before the assembly of the test battery. Ideally, tasks are used that are scalable in difficulty to maximize the variance in the dataset.

Finally, we suggest that it is time to go beyond *g* or the physical/social cognition divide. Executive functions with their strong genetic component [63], correlation to health markers [42] and domain generality (as established with humans) are arguably a prime candidate and a logical starting point for this endeavour. Measuring individual differences in EFs will also help to interpret individual variation in more specialized abilities. To date, most research in this area has been devoted to inhibitory control and we provide here evidence that the convergent validity of some widely used measures cannot be taken for granted and will require further investigation. Future experimental work is needed to establish reliable and valid measures of other EFs including attention shifting and working memory updating. Whenever possible, fitness and health measures and genetic samples might be added to the data collection to assess potential fitness consequences and to estimate heritability. In the long run, identifying the latent structure of cognitive abilities in a variety of species will allow us to trace back the evolutionary history of these abilities.

Data accessibility. Amici *et al*. [23] dataset: figshare (doi:10.6084/m9.figshare.5813904.v1) and MacLean *et al*. [24] dataset: figshare (doi:10.6084/m9.figshare.5579335.v1).

Authors' contributions. C.J.V. carried out the statistical analyses. All the authors helped draft the manuscript and gave their final approval for publication.

Competing interests. We declare we have no competing interests.

Funding. A.M.S. and C.J.V. were supported by the 'INQMINDS' ERC Starting Grant no. (SEP-210159400) awarded to A.M.S. B.T. was supported by a grant from the Social Sciences and Humanities Research Council of Canada (SSHRC 435-2016-1051).

Acknowledgements. We are grateful to Federica Amici and Evan MacLean for generously sharing their data with us.

References

1. Powell R, Mikhalevich I, Logan C, Clayton NS. 2017 Convergent minds: the evolution of cognitive complexity in nature. *Interface Focus* **7**, 20170029. (doi:10.1098/rsfs.2017.0029)
2. Dunbar RI. 1998 The social brain hypothesis. *Evol. Anthropol.* **6**, 178–190. (doi:10.1002/(SICI)1520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-8)
3. Barton RA. 1996 Neocortex size and behavioural ecology in primates. *Proc. R. Soc. Lond. B* **263**, 173–177. (doi:10.1098/rspb.1996.0028)
4. Ashton BJ, Thornton A, Ridley AR. 2018 An intraspecific appraisal of the social intelligence hypothesis. *Phil. Trans. R. Soc. B* **373**, 20170288. (doi:10.1098/rstb.2017.0288)
5. Milton K. 1981 Distribution patterns of tropical plant foods as an evolutionary stimulus to primate mental development. *Am. Anthropol.* **83**, 534–548. (doi:10.1525/aa.1981.83.3.02a00020)
6. Healy SD, Rowe C. 2007 A critique of comparative studies of brain size. *Proc. R. Soc. B* **274**, 453–464. (doi:10.1098/rspb.2006.3748)
7. Lefebvre L, Reader SM, Sol D. 2004 Brains, innovations and evolution in birds and primates. *Brain Behav. Evol.* **63**, 233–246. (doi:10.1159/000076784)
8. Reader S, Laland K. 2002 Social intelligence, innovation, and enhanced brain size in primates. *Proc. Natl Acad. Sci. USA* **99**, 4436–4441. (doi:10.1073/pnas.062041299)
9. Powell LE, Isler K, Barton RA. 2017 Re-evaluating the link between brain size and behavioural ecology in primates. *Proc. R. Soc. B* **284**, 20171765. (doi:10.1098/rspb.2017.1765)
10. Seed AM, Mayer C. 2017 Problem-solving in animals. In *APA handbook of comparative psychology* (eds J Call, GM Burghardt, IM Pepperberg, CT Snowdon, TR Zentall), pp. 601–625. Washington, DC: American Psychological Association.
11. Völter CJ, Call J. 2017 Causal and inferential reasoning in animals. In *APA handbook of comparative psychology: perception, learning, and cognition* (eds J Call, GM Burghardt, IM Pepperberg, CT Snowdon, TR Zentall), pp. 643–671. Washington, DC: American Psychological Association.
12. Mackintosh NJ. 1988 Approaches to the study of animal intelligence. *Br. J. Psychol.* **79**, 509–525. (doi:10.1111/j.2044-8295.1988.tb02749.x)

13. van Horik JO, Madden JR. 2016 A problem with problem solving: motivational traits, but not cognition, predict success on novel operant foraging tasks. *Anim. Behav.* **114**, 189–198. (doi:10.1016/j.anbehav.2016.02.006)
14. Thornton A, Lukas D. 2012 Individual variation in cognitive performance: developmental and evolutionary perspectives. *Phil. Trans. R. Soc. B* **367**, 2773–2783. (doi:10.1098/rstb.2012.0214)
15. Reader SM, Hager Y, Laland KN. 2011 The evolution of primate general and cultural intelligence. *Phil. Trans. R. Soc. B* **366**, 1017–1027. (doi:10.1098/rstb.2010.0342)
16. Herrmann E, Call J. 2012 Are there geniuses among the apes? *Phil. Trans. R. Soc. B* **367**, 2753–2761. (doi:10.1098/rstb.2012.0191)
17. Hebb DO, Williams K. 1946 A method of rating animal intelligence. *J. Gen. Psychol.* **34**, 59–65. (doi:10.1080/00221309.1946.10544520)
18. Burkart J, Schubiger M, van Schaik C. 2016 The evolution of general intelligence. *Behav. Brain Sci.* **40**, e195. (doi:10.1017/S0140525X16000959)
19. Borsboom D, Dolan CV. 2006 Why *g* is not an adaptation: a comment on Kanazawa (2004). *Psychol. Rev.* **113**, 433–437. (doi:10.1037/0033-295X.113.2.433)
20. Shuker DM, Barrett L, Dickins TE, Scott-Phillips TC, Barton RA. 2017 General intelligence does not help us understand cognitive evolution. *Behav. Brain Sci.* **40**, e218. (doi:10.1017/S0140525X16001771)
21. Van Der Maas HL, Dolan CV, Grasman RP, Wicherts JM, Huizenga HM, Raijmakers ME. 2006 A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychol. Rev.* **113**, 842. (doi:10.1037/0033-295X.113.4.842)
22. Cronbach LJ, Meehl PE. 1955 Construct validity in psychological tests. *Psychol. Bull.* **52**, 281. (doi:10.1037/h0040957)
23. Amici F, Aureli F, Call J. 2008 Fission-fusion dynamics, behavioral flexibility, and inhibitory control in primates. *Curr. Biol.* **18**, 1415–1419. (doi:10.1016/j.cub.2008.08.020)
24. MacLean EL *et al.* 2014 The evolution of self-control. *Proc. Natl Acad. Sci. USA* **111**, E2140–E2148. (doi:10.1073/pnas.1323533111)
25. Barth J, Call J. 2006 Tracking the displacement of objects: a series of tasks with great apes (*Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla*, and *Pongo pygmaeus*) and young children (*Homo sapiens*). *J. Exp. Psychol. Anim. Behav. Process.* **32**, 239–252. (doi:10.1037/0097-7403.32.3.239)
26. Donkers FC, Van Boxtel GJ. 2004 The N2 in go/no-go tasks reflects conflict monitoring not response inhibition. *Brain Cogn.* **56**, 165–176. (doi:10.1016/j.bandc.2004.04.005)
27. Seed AM, Seddon E, Greene B, Call J. 2012 Chimpanzee 'folk physics': bringing failures into focus. *Phil. Trans. R. Soc. B* **367**, 2743–2752. (doi:10.1098/rstb.2012.0222)
28. Campbell DT, Fiske DW. 1959 Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* **56**, 81. (doi:10.1037/h0046016)
29. Shaw RC, Schmelz M. 2017 Cognitive test batteries in animal cognition research: evaluating the past, present and future of comparative psychometrics. *Anim. Cogn.* **20**, 1003–1018. (doi:10.1007/s10071-017-1135-1)
30. Cole EF, Quinn JL. 2011 Personality and problem-solving performance explain competitive ability in the wild. *Proc. R. Soc. B* **279**, 20111539. (doi:10.1098/rspb.2011.1539)
31. Keagy J, Savard J-F, Borgia G. 2009 Male satin bowerbird problem-solving ability predicts mating success. *Anim. Behav.* **78**, 809–817. (doi:10.1016/j.anbehav.2009.07.011)
32. Huebner F, Fichtel C, Kappeler PM. 2018 Linking cognition with fitness in a wild primate: fitness correlates of problem-solving performance and spatial learning ability. *Phil. Trans. R. Soc. B* **373**, 20170295. (doi:10.198/rstb.2017.0295)
33. Isden J, Panayi C, Dingle C, Madden J. 2013 Performance in cognitive and problem-solving tasks in male spotted bowerbirds does not correlate with mating success. *Anim. Behav.* **86**, 829–838. (doi:10.1016/j.anbehav.2013.07.024)
34. Herrmann E, Call J, Hernández-Lloreda MV, Hare B, Tomasello M. 2007 Humans have evolved specialized skills of social cognition: the cultural intelligence hypothesis. *Science* **317**, 1360–1366. (doi:10.1126/science.1146282)
35. Lucon-Xiccato T, Bisazza A. 2017 Individual differences in cognition among teleost fishes. *Behav. Process.* **141**, 184–195. (doi:10.1016/j.beproc.2017.01.015)
36. Balakhonov D, Rose J. 2017 Crows rival monkeys in cognitive capacity. *Sci. Rep.* **7**, 8809. (doi:10.1038/s41598-017-09400-0)
37. Tomasello M, Call J. 1997 *Primate cognition*. New York, NY: Oxford University Press.
38. De Waal FB, Boesch C, Horner V, Whiten A. 2008 Comparing social skills of children and apes. *Science* **319**, 569. (doi:10.1126/science.319.5863.569c)
39. Herrmann E, Hare B, Cissewski J, Tomasello M. 2011 A comparison of temperament in nonhuman apes and human infants. *Dev. Sci.* **14**, 1393–1405. (doi:10.1111/j.1467-7687.2011.01082.x)
40. Herrmann E, Hernández-Lloreda MV, Call J, Hare B, Tomasello M. 2010 The structure of individual differences in the cognitive abilities of children and chimpanzees. *Psychol. Sci.* **21**, 102–110. (doi:10.1177/0956797609356511)
41. Hopkins WD, Russell JL, Schaeffer J. 2014 Chimpanzee intelligence is heritable. *Curr. Biol.* **24**, 1649–1652. (doi:10.1016/j.cub.2014.05.076)
42. Diamond A. 2013 Executive functions. *Annu. Rev. Psychol.* **64**, 135–168. (doi:10.1146/annurev-psych-113011-143750)
43. Friedman NP, Miyake A. 2017 Unity and diversity of executive functions: individual differences as a window on cognitive structure. *Cortex* **86**, 186–204. (doi:10.1016/j.cortex.2016.04.023)
44. Nigg JT. 2017 Annual Research Review: on the relations among self-regulation, self-control, executive functioning, effortful control, cognitive control, impulsivity, risk-taking, and inhibition for developmental psychopathology. *J. Child Psychol. Psychiatry* **58**, 361–383. (doi:10.1111/jcpp.12675)
45. MacLeod CM, Dodd MD, Sheard ED, Wilson DE, Bibi U. 2003 In opposition to inhibition. *Psychol. Learn. Motiv.* **43**, 163–215. (doi:10.1016/S0079-7421(03)01014-4)
46. Stahl C, Voss A, Schmitz F, Nuszbaum M, Tüscher O, Lieb K, Klauer KC. 2014 Behavioral components of impulsivity. *J. Exp. Psychol. Gen.* **143**, 850. (doi:10.1037/a0033981)
47. Friedman NP, Miyake A. 2004 The relations among inhibition and interference control functions: a latent-variable analysis. *J. Exp. Psychol. Gen.* **133**, 101–135. (doi:10.1037/0096-3445.133.1.101)
48. Hedge C, Powell G, Sumner P. 2017 The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods* **50**, 1166–1186. (doi:10.3758/s13428-017-0935-1)
49. Reynolds B, Ortengren A, Richards JB, de Wit H. 2006 Dimensions of impulsive behavior: personality and behavioral measures. *Personal. Individ. Diff.* **40**, 305–315. (doi:10.1016/j.paid.2005.03.024)
50. Aichert DS, Wöstmann NM, Costa A, Macare C, Wenig JR, Möller H-J, Rubia K, Ettinger U. 2012 Associations between trait impulsivity and prepotent response inhibition. *J. Clin. Exp. Neuropsychol.* **34**, 1016–1032. (doi:10.1080/13803395.2012.706261)
51. Duckworth AL, Kern ML. 2011 A meta-analysis of the convergent validity of self-control measures. *J. Res. Pers.* **45**, 259–268. (doi:10.1016/j.jrp.2011.02.004)
52. Beran MJ. 2015 The comparative science of 'self-control': what are we talking about? *Front. Psychol.* **6**, 51. (doi:10.3389/fpsyg.2015.00051)
53. Bari A, Robbins TW. 2013 Inhibition and impulsivity: behavioral and neural basis of response control. *Prog. Neurobiol.* **108**, 44–79. (doi:10.1016/j.pneurobio.2013.06.005)
54. Verbruggen F, McLaren IP, Chambers CD. 2014 Banishing the control homunculi in studies of action control and behavior change. *Perspect. Psychol. Sci.* **9**, 497–524. (doi:10.1177/1745691614526414)
55. Kabadayi C, Taylor LA, von Bayern AM, Osvath M. 2016 Ravens, New Caledonian crows and jackdaws parallel great apes in motor self-regulation despite smaller brains. *R. Soc. open sci.* **3**, 160104. (doi:10.1098/rsos.160104)
56. Jelbert S, Taylor A, Gray R. 2016 Does absolute brain size really predict self-control? Hand-tracking training improves performance on the A-not-B task. *Biol. Lett.* **12**, 20150871. (doi:10.1098/rsbl.2015.0871)
57. Bray EE, MacLean EL, Hare BA. 2014 Context specificity of inhibitory control in dogs. *Anim. Cogn.* **17**, 15–31. (doi:10.1007/s10071-013-0633-z)
58. van Horik JO, Langley EJ, Whiteside MA, Laker PR, Beardsworth CE, Madden JR. 2018 Do detour tasks provide accurate assays of inhibitory control? *Proc. R. Soc. B* **285**, 20180150. (doi:10.1098/rspb.2018.0150)

59. Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD. 2000 The unity and diversity of executive functions and their contributions to complex 'frontal lobe' tasks: a latent variable analysis. *Cogn. Psychol.* **41**, 49–100. (doi:10.1006/cogp.1999.0734)
60. Dougherty LR, Guillette LM. 2018 Linking personality and cognition: a meta-analysis. *Phil. Trans. R. Soc. B* **373**, 20170282. (doi:10.1098/rstb.2017.0282)
61. Sih A, Del Giudice M. 2012 Linking behavioural syndromes and cognition: a behavioural ecology perspective. *Phil. Trans. R. Soc. B* **367**, 2762–2772. (doi:10.1098/rstb.2012.0216)
62. Cronbach LJ, Furby L. 1970 How we should measure 'change': Or should we? *Psychol. Bull.* **74**, 68. (doi:10.1037/h0029382)
63. Friedman NP, Miyake A, Young SE, DeFries JC, Corley RP, Hewitt JK. 2008 Individual differences in executive functions are almost entirely genetic in origin. *J. Exp. Psychol. Gen.* **137**, 201–225. (doi:10.1037/0096-3445.137.2.201)
64. Hall PA, Crossley M, D'Arcy C. 2010 Executive function and survival in the context of chronic illness. *Ann. Behav. Med.* **39**, 119–127. (doi:10.1007/s12160-010-9162-z)
65. Miyake A, Friedman NP. 2012 The nature and organization of individual differences in executive functions: four general conclusions. *Curr. Dir. Psychol. Sci.* **21**, 8–14. (doi:10.1177/0963721411429458)
66. Wiebe SA, Sheffield T, Nelson JM, Clark CA, Chevalier N, Espy KA. 2011 The structure of executive function in 3-year-olds. *J. Exp. Child Psychol.* **108**, 436–452. (doi:10.1016/j.jecp.2010.08.008)
67. Wiebe SA, Espy KA, Charak D. 2008 Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure. *Dev. Psychol.* **44**, 575. (doi:10.1037/0012-1649.44.2.575)
68. Bechtel W, Richardson RC. 2010 *Discovering complexity: decomposition and localization as strategies in scientific research*. Cambridge, MA: MIT Press.
69. Botvinick MM, Cohen JD. 2014 The computational and neural basis of cognitive control: charted territory and new frontiers. *Cogn. Sci.* **38**, 1249–1285. (doi:10.1111/cogs.12126)
70. Rougier NP, Noelle DC, Braver TS, Cohen JD, O'Reilly RC. 2005 Prefrontal cortex and flexible cognitive control: rules without symbols. *Proc. Natl Acad. Sci. USA* **102**, 7338–7343. (doi:10.1073/pnas.0502455102)
71. Carruthers P. 2013 Evolution of working memory. *Proc. Natl Acad. Sci. USA* **110**, 10 371–10 378. (doi:10.1073/pnas.1301195110)
72. Kolata S, Light K, Grossman HC, Hale G, Matzel LD. 2007 Selective attention is a primary determinant of the relationship between working memory and general learning ability in outbred mice. *Learn. Mem.* **14**, 22–28. (doi:10.1101/lm.408507)
73. Kolata S, Light K, Townsend DA, Hale G, Grossman HC, Matzel LD. 2005 Variations in working memory capacity predict individual differences in general learning abilities among genetically diverse mice. *Neurobiol. Learn. Mem.* **84**, 241–246. (doi:10.1016/j.nlm.2005.07.006)
74. Cauchoix M *et al.* 2018 The repeatability of cognitive performance: a meta-analysis. *Phil. Trans. R. Soc. B* **373**, 20170281. (doi.10.1098/rstb.2017.0281)
75. Mayer CP. 2015 *The evolutionary origins of executive functions: behavioural control in humans and chimpanzees*. St Andrews, UK: University of St Andrews.
76. Petrides M. 1995 Impairments on nonspatial self-ordered and externally ordered working memory tasks after lesions of the mid-dorsal part of the lateral frontal cortex in the monkey. *J. Neurosci.* **15**, 359–375. (doi:10.1523/JNEUROSCI.15-01-00359.1995)
77. Washburn DA. 1994 Stroop-like effects for monkeys and humans: processing speed or strength of association? *Psychol. Sci.* **5**, 375–379. (doi:10.1111/j.1467-9280.1994.tb00288.x)
78. Middlebrooks PG, Schall JD. 2014 Response inhibition during perceptual decision making in humans and macaques. *Atten. Percept Psychophys* **76**, 353–366. (doi:10.3758/s13414-013-0599-6)
79. Bargh JA, Ferguson MJ. 2000 Beyond behaviorism: on the automaticity of higher mental processes. *Psychol. Bull.* **126**, 925. (doi:10.1037/0033-2909.126.6.925)
80. Baker ST, Gjersoe NL, Sibielska-Woch K, Leslie AM, Hood BM. 2011 Inhibitory control interacts with core knowledge in toddlers' manual search for an occluded object. *Dev. Sci.* **14**, 270–279. (doi:10.1111/j.1467-7687.2010.00972.x)
81. Benson JE, Sabbagh MA, Carlson SM, Zelazo PD. 2013 Individual differences in executive functioning predict preschoolers' improvement from theory-of-mind training. *Dev. Psychol.* **49**, 1615. (doi:10.1037/a0031056)
82. Coolidge FL, Wynn T. 2001 Executive functions of the frontal lobes and the evolutionary ascendancy of *Homo sapiens*. *Camb. Archaeol. J.* **11**, 255–260. (doi:10.1017/S0959774301000142)
83. Frank MC. 2015 The ManyBabies Project. See https://manybabies.github.io/ (accessed 12 July 2018).
84. Bohn M, Schmitt V, Sanchez-Amaro A, Keupp S, Hopper L, Völter C, Altschul D, Fischer J, Fichtel C. 2018 ManyPrimates. *Open Sci., Framework* 2018. See osf.io/v5je6 (accessed 12 July 2018).