A preclinical cognitive test battery to parallel the National Institute of Health Toolbox in humans: bridging the translational gap

Shikha Snigdha a,*, Norton W. Milgram b, Sherry L. Willis c, Marylin Albert d, S. Weintraub e, Norbert J. Fortin f,g, Carl W. Cotman a,g

a Institute for Memory Impairments and Neurological Disorders, University of California—Irvine, Irvine, California, USA
b CanCog Technologies Inc, Toronto, Ontario, Canada
c Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA
d Division of Cognitive Neuroscience in the Department of Neurology at Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
e Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA
f Center for the Neurobiology of Learning and Memory, University of California—Irvine, Irvine, California, USA
g Department of Neurobiology and Behavior, University of California—Irvine, Irvine, California, USA

1. Introduction

There has been little uniformity among measures used in human neuropsychological assessment to assess cognitive deficits associated with aging and Alzheimer’s disease (AD). This has had a negative impact on the pace of discovery in research on aging and dementia in clinical trials, and confounds the utility of preclinical assessment. Accordingly, the National Institutes of Health supported the development of a comprehensive assessment tool, the NIH Toolbox For Assessment Of Neurological And Behavioral Function, for use in longitudinal, epidemiological, and intervention studies. The entire NIH Toolbox covers emotion, cognition, motor, and sensory function, whereas the cognition section of the NIH Toolbox provides a specific neuropsychological instrument battery (NIH Toolbox Cognitive Function Battery [CFB]), to probe several cognitive domains (working memory, episodic memory, attention, executive function, processing speed, language, and reading). To move the field of translational research forward, there is an urgent need to develop a comparable preclinical test battery using animal models.

Use of animal models for evaluation of cognitive dysfunction involves simulating specific behaviors or symptoms associated with human cognition. Three important validation criteria for evaluation of such model systems are face, construct, and predictive validity. Although construct validity relies on a match between the proposed pathophysiology of a condition and that of the model species, predictive validity focuses on a match with clinical studies in its response to interventions. The third criterion used to define validity of a model is “face validity,” which relies on a match between the behavioral effects observed in a model and those exhibited by the species being modeled. Although it is used in initial design of tests, it is a less stringent criterion compared with construct and predictive validity. Two types of models that are traditionally used in brain aging and AD research are discussed here.
11. Natural or spontaneous models

These include species that show a natural deposition of amyloid and tau proteins, both hallmarks of the disease condition in humans, together with cognitive decline. Nonhuman primates (Gearing et al., 1997; Martin et al., 1994; Price et al., 1991) and dogs (Cotman and Head, 2008; Head and Torp, 2002; Pugliese et al., 2006) are 2 of the most common species used as spontaneous models. These animals have a well-developed prefrontal cortex and a relatively long lifespan. This ensures that the animals can perform higher-order cognitive tasks and exhibit age-induced behavioral abnormalities that parallel cognitive deficits shown by aged humans.

The rhesus monkey, for instance, has a lifespan of more than 30 years. This is equivalent to about 90 human years (Price et al., 1991). These monkeys share a 92% to 95% genetic homology with humans and, similar to humans, show age-related cognitive impairments relative to their young counterparts (Herndon et al., 1997; Smith et al., 2004). Aged monkeys also show loss of cholinergic activity (as seen in human AD patients) and deposition of amyloid plaques that can be visualized by using human anti-Aβ protein antibodies (Summers et al., 1997; Voytko et al., 2001). Furthermore, pharmacological interventions that increase acetylcholine release have been successfully shown to enhance cognitive performance in aged monkeys (Terry et al., 1993).

The aged dog, which is considered to be 1 of the best and most accessible animal models of brain aging, also shows many of the key features of human brain aging, mild cognitive impairment (MCI), and early AD (Cotman and Head, 2008; Sarasa and Pesini, 2009). Like the aged human brain, the canine brain shows increased oxidative damage, mitochondrial dysfunction, selective neuron loss, decreased hippocampal neurogenesis, and accumulation of beta-amyloid (Aβ) pathology with age. The aged canine is a natural model of Aβ accumulation, as the canine and human Aβ protein is 100% homologous and the APP sequences share 98% homology. Indeed, the canine brain accumulates senile plaques with age, and the accumulation of Aβ1-42 and Aβ1-40 progresses in a similar fashion to that occurring in the human brain. Furthermore, pharmacological and dietary interventions, and exercise have been shown to improve health and cognitive decline in the aged dogs, lending some predictive validity to the use of the model (Cotman and Head, 2008; Fahnestock et al., 2012; Milgram et al., 2005).

However, despite the significant advantages of higher-animal models, rodents in general, and rats and mice in particular, continue to remain the most commonly used animals as experimental models of aging and AD. A high birth rate, short reproductive and life cycle, and small size make them ideal laboratory animals. Furthermore, with the increased use of transgenic and knockout mouse models, it has become more common to use induced models to evaluate the etiology of the disease and to develop pharmaceutical treatments.

12. Induced models

These models rely on induction of AD-like pathology using experimental manipulations such as lesions (Lescaudron and Stein, 1999 Vale-Martinez et al., 2002), drugs (Buccafusco, 2009; Decker and McGaugh, 1991; Taffe et al., 1999), amyloid-β infusion and genetic alterations (for review see (Gotz and Ittner, 2008; Zahs and Ashe, 2010)). Drug treatments that produce a disruption or loss of acetylcholine brain transmission are examples of pharmacological disruption that models Alzheimer’s dementia. Such a deficit can be reproduced in animals by blocking cholinergic receptors pharmacologically with drugs such as scopolamine and mecamylamine or by using neurotoxic, electrolytic, or mechanical lesions of cholinergic brain subregions (Toledano et al., 2010). Lesion models have the advantage of producing more chronic deficits than the pharmacological models. Genetic manipulations include knockouts, knockdowns, and transgenic mice, and are extensively used to study pathological characteristics of AD, such as amyloid plaques and neurofibrillary tangles. Despite some drawbacks (reviewed in Zahs and Ashe, 2010), genetically altered species can be considered the most valid induced model based on similarities in construct (Zahs and Ashe, 2010).

Although the study of aging and AD depends on the use of animal models, the correspondence between the behavioral assays used in humans and animals to measure the key cognitive domains of the NIH Toolbox is not well established. In the following sections, we describe and evaluate tests that can be used to measure subdomains from the NIH Toolbox CFB (overview in Table 1). Such a test battery would facilitate study comparisons between different groups, make data pooling much more feasible, and improve the translational properties of preclinical research. Our goal here is to outline a battery of preclinical behavioral tests that may be used to assess 5 of the 7 cognitive domains comprising the NIH Toolbox CFB. Language and reading, which involve drawing inferences from written or printed text, are extremely difficult to reliably model in animals, and will not be addressed in this article.

### Table 1: Overview of animal tasks that parallel the NIH Toolbox cognitive function battery

<table>
<thead>
<tr>
<th>Domain</th>
<th>Task in NIH Toolbox CFB</th>
<th>Parallel animal task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive flexibility</td>
<td>Dimensional card sort task</td>
<td>Attentional set-shifting task</td>
</tr>
<tr>
<td>Inhibitory control</td>
<td>Flanker task</td>
<td>Reversal learning task</td>
</tr>
<tr>
<td>Working memory</td>
<td>Single and multiple list sorting task</td>
<td>Go/No-go tasks</td>
</tr>
<tr>
<td>Episodic memory</td>
<td>Sequential memory/learning task</td>
<td>Discussed as a separate domain</td>
</tr>
<tr>
<td>Working memory</td>
<td>Single and multiple list sorting task</td>
<td>What—what tasks</td>
</tr>
<tr>
<td>Processing speed</td>
<td>Pattern comparison task</td>
<td>DNMP/DNMS</td>
</tr>
<tr>
<td>Language</td>
<td>Picture vocabulary test</td>
<td>Self-order tasks</td>
</tr>
</tbody>
</table>

Key: CFB, cognitive function battery; DNMP/DNMS, delayed non-match to position/delayed non-match to sample; N/A, not available; NIH, National Institutes of Health.

### 2. NIH Toolbox CFB and comparable preclinical tests

2.1. Executive function

Overall, the NIH Toolbox CFB considers executive function as the top-down cognitive modulation of goal-directed activity. Executive function involves several different components or subdomains, including set shifting, inhibitory response, and updating/working memory. In the NIH Toolbox CFB, the Dimensional Change Card Sort task is used as a measure of set shifting. This task is a measure that was adapted for adults from pediatric research (Zelazo, 2006). In this task, subjects must respond to pairs of stimuli by selecting the one that is the same shape as a third target stimulus or one that is the same color. The sorting contingencies shift, requiring the ability to inhibit previous response strategies and to try new ones. In accordance, we propose that the attentional set-shifting task be used as a measure of set-shifting in animals.

The NIH Toolbox CFB describes a flanker task to evaluate the inhibitory component of visual attention and executive function. The test presents a line of arrows (or fish, for children) in the center of the computer screen (Fan et al., 2002). The central stimulus points leftward or rightward, whereas the others point either in the same...
direction as the central stimuli (congruent trials) or in the opposite direction (incongruent trials). The task is to focus on the central arrow and indicate its direction, ignoring the “flankers.” Reaction time is slower on incongruent than on congruent trials, reflecting the contribution of an inhibitory response to the incongruent flankers. This task on the NIH Toolbox CFB yields a score for the executive component and can also yield a score reflecting responses to congruent conditions, reflecting in turn sustained attention without inhibitory responses. Here, we propose that Go/No-go paradigms can provide excellent tests to measure inhibitory response.

Working memory, which is both a component of executive function and a separate domain in the NIH Toolbox, will be discussed later here.

2.1. Attentional set shifting

Cognitive flexibility in animals can be measured by using an attentional set-shifting or intra-extra-dimensional shift task (Birrell and Brown, 2000; Dias et al., 1996). In this test, animals are trained to carry out a series of 7 discriminations (Fig. 1A), namely, simple discrimination (SD), compound discrimination (CD), reversal 1 (R1), intra-dimensional shift (IDS), reversal 2 (R2), extra-dimensional shift (EDS), and reversal 3 (R3). Animals are trained to select the target (associated with the positive reward) based on a single criterion. For instance, in rodent studies, the first stage is often an SD, in which the animals chose 1 food bowl over the other based on the shape of bowl. This is identical to the SD in the training session on the previous day, except that new bowls are used. The animal is expected to reach a predetermined criterion of “n” consecutive correct responses. For CD, a second dimension (odor) is introduced, but the correct and incorrect exemplars remain the same (shape of bowl). For the IDS, the relevant stimulus is still the shape of bowl, but a new shape is introduced. For the reversal phase, the exemplars and relevant dimensions remain the same, but the animals have to learn that the previously baited shape or odor is now incorrect and the other odor is now the correct one. For the ED shift, the previously irrelevant parameter (i.e., odor) is now relevant. It should be noted that new exemplars are used for the ID and ED shifts.

2.1.2. Task validity and effects of aging

Manipulations, which impair performance in the set shifting task, include lesions in the medial prefrontal cortex (PFC) in rats (Birrell and Brown, 2000; Dias et al., 1996) and dorsolateral PFC in monkeys (Dias et al., 1996). Aged rats are also impaired in task performance (Nicolle and Baxter, 2003). Pharmacologically, scopolamine selectively disrupts task performance during the ED shift but not in the ID phase (Chen et al., 2004). This dissociation is consistent with human studies using the Wisconsin Card Sorting Test (WCST) and is suggestive of some construct validity (Owen et al., 1991). Fewer studies have used mice to evaluate set-shifting deficits. Recently, Young et al. have used the set-shifting task in mice to show that aged mice may use a strategy of “speed to accuracy trade off” to maintain accurate performance in the task (Young et al., 2010). Specifically, these investigators reported that older mice were not specifically impaired in trials to criterion but were significantly slower in making correct responses. The only other study that used the attentional set-shifting task in a murine model used a transgenic APP precursor protein mutation model (TG2576). The authors reported an overall effect of aging, but there was no evidence of an ID/ED shift or performance difference between simple and reversal phases, which in turn questions the formation of attentional set in the mice (Zhuo et al., 2007). A key difference between the 2 studies was the use of normal aged versus transgenic mice. Collectively, these studies suggest that the attentional set-shifting task is a good model of set shifting in primates and rats, but that additional work is needed to adapt it to mice.

Reversal learning, which is often used as an independent test of executive function, is a part of the set-shifting task described previously. However, when used as an independent test, rodent models commonly use an operant chamber (Neese et al., 2010) or odor discrimination/reversal task (Schoenbaum et al., 2002a) to measure reversal learning. Briefly, the operant chamber method uses a 2-lever operant box to first train the animals on 1 lever and not the other, following which the active lever is switched, requiring the rat to reverse its response to the previously nonbaited lever. This is similar to the R1 phase of the attentional set-shifting task. However, a study by Roberts and Wallis (2000) in primates, showed that whereas set-shifting ability is impaired by lateral lesions of the PFC, reversal learning is impaired by orbital but not lateral lesions. This is consistent with the work of Schoenbaum et al. (Schoenbaum et al., 2003) in rodents. It must be noted that Schoenbaum have also reported that the orbitofrontal cortex is necessary for inhibiting responses only when the responses need to be altered to reflect changing relationships between cues and outcomes (Schoenbaum et al., 2002b). This suggests that a loss of inhibitory control can occur independently from other components of executive function. Although reversal learning can be used to test cognitive flexibility and inhibitory control, it is difficult to isolate these 2 components in the task. Thus, we recommend investigating set shifting and inhibitory control independently using the attentional set-shifting task and go/no-go paradigms, respectively.

2.1.3. Go/No-go paradigms

A go/no-go task requires subjects to perform a learned response on “go” trials (e.g., to push a lever in response to a color) but withholding responding on “no-go” trials (e.g., not to push lever on a different color) (Fig. 1B). The index of inhibitory control is the number of errors that a subject makes on no-go trials (Iversen and Mishkin, 1970, Sakagami and Tsutsui, 1999, Sakagami et al., 2001).

2.1.4. Task validity and effects of aging

The main function of the PFC is one of “executive control”; however, there is some controversy over whether subregions of PFC are functionally differentiated in executing this control, and whether there is localization of inhibitory functions to dorsolateral PFC (Diamond, 1990), orbital frontal cortex or right inferior frontal cortex (IFC) (Aron et al., 2004; Roberts and Wallis, 2000). Nonetheless most studies agree that the prefrontal regions of the brain are involved in inhibitory responses and in signaling for overriding an automatic response tendency. Neuroimaging studies using go/no-go tasks consistently point to involvement of a right-lateralized IFC region (Garavan et al., 2002; Menon et al., 2001; Nielson et al., 2002). This has also been demonstrated in nonhuman primate studies, lending further credibility to the model (Iversen and Mishkin, 1970). The differences in the specific subregions underlying control of different tasks listed here suggest that the ability to suppress response tendencies may be an intrinsic property of the prefrontal cortex, whereas the form of the “disinhibition” is more dependent on the nature of the whole operation (or task) being performed.

2.2. Episodic memory

Episodic memory refers to the ability to remember information about personal experiences along with the context in which they occurred (Dickerson and Eichenbaum, 2010; Tulving and Thomson, 1973; Tulving, 1983). This specific form of memory is the last to be acquired during childhood (Perner and Ruffman, 1995) and is among the first to be lost with progressive aging (Herlitz and Forsell, 1996). The NIH Toolbox CFB refers to episodic memory as the sum of cognitive processes involved in the acquisition, storage,
and effortful recall of episodes or information within a context. The task incorporated into the NIH Toolbox CFB is the Picture Sequence Memory Test. It was developed for adults from the Imitation Based Assessment of Memory paradigm used to test learning and recall in infants and young children (Bauer, 1993). Single pictures depicting scenes with a common theme (e.g., playing in the park) are sequentially presented, after which the pictures are “gathered” on the computer screen to the center. The subject must recapitulate

Fig. 1. (A) Attentional set-shifting task. The image shows a common form of the attentional set-shifting task. Rats are trained to choose the appropriate bowl to dig for food reward based on defined discrimination rules that change at every successive stage of the task. Each bowl can vary in the shape of the bowl and the odor of the media. Intradimensional shift comprises a total stimulus change, but with the same measure (odor or shape) still being the relevant stimulus to attend to. During extradimensional shifting, the other dimension, which had so far been irrelevant, becomes the relevant measure. (B) Go/No-go task. Representative image showing subjects performing on go trials (in response to the color green on screen) and inhibiting response on no-go trials (in response to the color red on screen).
the sequence demonstrated by the examiner. There are 3 trials, and
the sequence length varies depending on the age group being
tested: 6 pictures for 3- to 4-year-olds; 9 pictures for 5- to 6-year-
olds; 12 for 7- and 8-year-olds; 15 for subjects aged 9 to 60 years;
and 9 pictures for subjects aged 65 to 85 years. Because of
constraints on the total time allocated to the NIH Toolbox CFB,
there is no delayed recall trial. However, there is ample evidence to
support a very high correlation between performance on learning
trials and subsequent recall of the same material after a delay.

As mentioned earlier, a defining feature of episodic memory is
that the memory for individual events is intrinsically tied to the
context in which they occurred. This operational definition is the
basis of most animal models of episodic memory. In most para-
digms, animals demonstrate episodic memory capacity by re-
membering specific events as well as the spatial and/or temporal
context in which they occurred. Approaches developed to study
episodic memory in animals range from the original model of
memory for “what—where—when” (Clayton and Dickinson, 1999) to
more recent characterizations of the retrieval dynamics of episodic
recollection using receiver-operating characteristics (ROC) (Fortin
et al., 2004; Sauvage et al., 2008). These approaches are compre-
hensively reviewed elsewhere (Clayton et al., 2003; Crystal, 2009;
Dere et al., 2006; Eichenbaum et al., 2012; Fouquet et al., 2010;
Kesner and Hunsaker, 2010). Here, we consider approaches focusing
on the spatial context, the temporal context, and the integration of
the 2 contexts.

2.2.1. Memory for events in their spatial context (what—where)

The common cognitive demand of the following tasks is to
remember where a specific event (usually the presentation of a
specific stimulus) occurred in the environment (Fig. 2A). This
capacity involves distinguishing similar events and/or spatial
locations and is thought to depend on pattern separation processes
(Gilbert et al., 1998; Yassa et al., 2011b). The “where” component
can refer to a specific place in an environment, or to a specific
location on a screen or complex visual scene.

This is in contrast to simple left/right position differences
between stimuli as seen in delayed nonmatch or match to sample
tasks. In rodents, a common approach is to use paired-associates
paradigms in which animals must remember specific odor—place
(Day et al., 2003, Rajji et al., 2006) or object—place associations
(Gilbert and Kesner, 2003, 2004). Similar item-in-place paradigms
have been developed using computer screens in nonhuman
primates (Baxter et al., 2007; Gaffan, 1994), and have proved
successful in detecting early symptoms of cognitive decline associ-
ated with AD dementia (Taffe et al., 2004). Overall, these approaches
can also be adapted to evaluate episodic memory in other animals
such as rodents, dogs, and nonhuman primates.

Fig. 2. (A) One general approach for paradigms that aim to evaluate episodic memory is that animals demonstrate episodic memory capacity by remembering specific events as well as the spatial context in which they occurred. The image shows the principle underlying cognitive demands involved in spatial episodic memory tasks, which is to remember the association between presentation of a specific stimulus (oranges) and the specific environment (grass) in which it occurred. (B) Memory for events in their temporal context: In each trial, rats are presented with a series of odors, following which the animal is probed for its memory of the order of elements in the series. The stimulus (odor) associated with reward in the retention phase is the stimulus that was presented first in the sample phase.
2.2. Task validity and effects of aging

The main advantages of this general approach are that the tasks are relatively simple to administer, use common apparatus, and are learned relatively quickly. The latter is particularly true for tasks relying on spontaneous preference measures (Dix and Aggleton, 1999; Langston and Wood, 2010; Lee et al., 2005; Mumbry et al., 2002a; Mumbry et al., 2002b). Overall these paradigms have proved to be sensitive measures of hippocampal dysfunction (Kesner et al., 2008) and of episodic memory dysfunction in the preclinical stages of AD (Fowler et al., 2002; Lee et al., 2003; Swainson et al., 2001). In particular, changes in the CA3 and dentate gyrus of the hippocampus are implicated in decreased efficiency in pattern separation in the aging brain (Holden et al., 2012; Vassa et al., 2011a). This measure of episodic memory is a considerable improvement over item-based recognition tests (Dalla Barba, 1997; Lipinska and Backman, 1997; Tierney et al., 1996; Welsh et al., 1991), which rely more on recognition memory than episodic memory. Finally, it should be noted that spatial paradigms do not strongly parallel the NIH Toolbox CFB measure of episodic memory, which is a nonspatial sequence memory task.

2.2.3. Memory for events in their temporal context (what—when)

This approach focuses on the memory for “when” events occurred. Importantly, this approach is nonspatial, so the memory for the temporal context can be studied independently of the spatial context. One task of particular interest has focused on the capacity to remember trial-unique sequences of events (Fortin et al., 2002; Kesner et al., 2002). In this task, animals are presented with a series of randomly selected odors in the sample phase. A single-choice test is then used to probe for memory of the sequence where the animal is rewarded for selecting the odor that had appeared earlier in the series (Fig. 2B). The task also includes a control condition in which animals are tested for their memory (recognition) of the odors (but not the temporal relationship), a capacity that can be supported by recognition memory processes. Animals with damage to the hippocampus can remember the individual odors presented in each sequence, but not their temporal, sequence-specific relationship (Fortin et al., 2002; Kesner et al., 2002). Similar tasks have been used in monkeys (Kumaran and Maguire, 2006a; Naya and Suzuki, 2011; Petrides, 1995) and humans (Kumaran and Maguire, 2006a; 2006b) as well.

2.2.4. Task validity and effects of aging

The main advantage of this general approach is that it closely parallels the Picture Sequence Memory Test used to measure episodic memory in the Toolbox, lending much face validity to the model. The main disadvantage is that it requires extensive training to teach animals the rules of the task and to comprehensively assess sequence memory capacity. Note that other types of sequence memory tasks have been developed, including versions based on spontaneous preference (Hauser et al., 2009; Hoge and Kesner, 2007), which provide a simpler, but quicker, assay of the capacity to remember sequences of events, as reviewed by Kesner and Churchwell (Kesner and Churchwell, 2011). Furthermore, the hippocampus has been shown to be essential for successful execution of the sequence memory task (Fortin et al., 2002; Kesner et al., 2002) and for animals to retain the ability to disambiguate overlapping sequences of odors (Agster et al., 2002). However, to the best of our knowledge, no animal studies have attempted to directly use the sequence memory task mentioned above to evaluate episodic memory in aging, making it difficult to comment on the predictive validity of the task.

2.2.5. “What—where—when” paradigms

The third approach is centered on the original animal model of episodic memory, the memory for what—where—when in food-storing scrub jays (Clayton and Dickinson, 1999). Whereas the studies described earlier focused on either the spatial or the temporal context, this approach requires memory for both types of contexts.

The original model took advantage of the natural caching behavior of scrub jays, to demonstrate that they could remember specific episodic memories: they could remember “what” they stored (worms or peanuts), and well as “where” (the location in the cage) and “when” (4 hours or 124 hours ago) they did it. Since then, what—where—when memory has been demonstrated in rodents (Babb and Crystal, 2006a, b; Eacott and Norman, 2004; Ergorul and Eichenbaum, 2004; Kart-Teke et al., 2006) and primates (Hoffman et al., 2008).

2.2.6. Task validity and effects of aging

A positive aspect of this approach is that it has a very stringent criterion for the demonstration of episodic memory in animals. In that sense, this model could be considered better than the others. However, the stringent criterion is also a downside, as simultaneously probing for what, where, and when makes it difficult to isolate the nature of memory impairments or the critical neural substrates. For that reason, approaches focusing on either the spatial or temporal context appear to be better suited for investigating the effects of aging or AD. It should be noted that this approach also requires extensive training, although simpler versions using spontaneous preference tests have also been developed (Kart-Teke et al., 2006).

2.3. Working memory

As defined in the NIH Toolbox CFB, working memory is a step up from the conventional construct of short-term memory storage and clearly includes the concept of an “active computational workspace.” The task, based on one developed by Mungas et al. for English and Spanish speakers (Mungas et al., 2011), consists of sequential presentation of pictures of objects and their corresponding auditory words (e.g., “strawberry,” “apple”). Once the series, which varies in the number of items presented, is complete, the subject must report the objects in order of size. In 1 condition, all items from a sequence are from a single category (e.g., animals or fruit). In a second condition, items from 2 categories are presented in random sequence, and subjects are instructed to report the items in order of size from smallest to largest but also to list all items from 1 category first and then all items from the second. Therefore, for a comparable animal test paradigm it is essential to include preclinical tests that require short-term storage of information together with manipulation of information to achieve a defined goal. Unfortunately, current animal models of working memory do not fully capture both features. Here, we highlight delay-type tasks that are most suitable for examining the short-term memory storage aspect of working memory, and we identify promising approaches for investigating the capacity of animals to manipulate information in working memory.

2.3.1. Short-term memory storage

There is a large body of literature showing that primates and rodents can hold specific stimuli “in mind” for short periods of time (reviewed in Castner et al., 2004; Dudchenko et al., 2012). Typically, this capacity is investigated using delay-type paradigms, in which each trial is divided in 3 phases: (1) a sample phase, in which the animal is presented with the stimulus to remember; (2) a delay phase, during which the stimulus is not present; and (3) a choice phase in which the animal is rewarded for demonstrating its memory of the sample stimulus. Many variants of this general approach have been developed, including the use of different types...
of sample stimuli (e.g., pictures, objects, tones, levers, locations), reward contingencies for the choice phase (“matching” or “non-matching” to the sample), and apparatus (manual or automated). Overall, these different versions provide consistent findings (Dudchenko et al., 2012); however, before choosing a specific version, it is important to consider the following. First, it is critical that sample stimuli be small and repeated across trials. This interference is important to ensure that animals cannot solve the task based on the degree of familiarity with particular stimuli (recognition memory), or by storing and subsequently retrieving the specific experience of being presented with the stimulus (episodic memory). Second, as discussed in Jenesson and Squire (2012), it is preferable not to test short-term memory in rodents using tasks requiring complex spatial processing (e.g., use specific arm entries as sample stimuli). Not only is the use of complex spatial cues not typical in human (Stern et al., 2001) and nonhuman primate studies (Carlson et al., 1997; Shamy et al., 2011), it increases the likelihood that performance will critically depend on the hippocampus (Olton and Papas, 1979; Olton et al., 1978), which is generally not the case for the conventional version of the task involving simpler stimuli (e.g., visual or auditory stimuli, position of levers or items on a screen). Third, although the use of either match or non-match contingencies lead to similar findings (Dudchenko et al., 2012), animals tend to learn the rules of the tasks more quickly if a non-match contingency is used (Mishkin and Delacour, 1975).

For these reasons, we recommend the use of the Delayed Non-Match to Position task (DNMP) (Bussey et al., 2003; Dunnett et al., 1988) to study the short-term memory storage component of working memory. This task has been used extensively in rodents, so considerable data are available on its critical neural substrate and pharmacology (reviewed by Dudchenko et al., 2012). Similar paradigms have been widely used in nonhuman primate studies as well (delayed-response task, Funahashi et al., 1989; Joseph and Barone, 1987). The DNMP task can be run manually or with an automated apparatus (with levers, or lights on touchscreen), the latter being more advantageous in terms of standardization and throughput. The Delayed Non-Match to Sample Task (DNMS) (Mishkin and Delacour, 1975; Rapp and Amaral, 1989) is a good alternative to the DNMP; it is essentially the same task, with the exception that sample stimuli are discrete cues (e.g., pictures, objects, tones) rather than positions. However, as mentioned earlier, it is critical that sample stimuli be repeated to avoid tapping into recognition memory processes.

2.3.2. Manipulation of information in working memory

In addition to the short-term storage of information, the NIH Toolbox definition of working memory includes a requirement for “on line” manipulation of that information. To our knowledge, the type of manipulation required by the Mungas et al. (2011) task has not been demonstrated in animals; however, simpler forms have been reported in nonhuman primates, such as the capacity to “monitor” information within working memory. For instance, monkeys can keep track of the specific sequence (Fig. 3) in which highly familiar stimuli were last presented (Petrides, 1991a; Warden and Miller, 2010), and can keep track of which stimuli from a similar set have already been rewarded on the current trials (Petrides, 1995). This capacity has been shown to depend on the integrity of the dorsolateral prefrontal cortex in monkeys as well as in humans (Petrides, 1991a, b, Petrides et al., 1993a, Petrides et al., 1993b, Petrides, 1995). Although rodents can remember trial-unique sequences of stimuli (Fortin et al., 2002; Kesner et al., 2002), their ability to do so using a set of repeated stimuli has not been well characterized.

Other forms of monitoring of information in working memory include the capacity to perform “n-back” tasks (Barch et al., 1997; Cohen et al., 1997), which require subjects to indicate whether the current stimulus matches the one n steps earlier in the series. Older human subjects have been shown to perform as well as younger subjects at 1-Back version of this task but are significantly impaired at 2-Back and 3-Back versions (Matty et al., 2006). The n-back task in humans is a continuous performance task in which every stimulus that is presented is used as a cue to guide the next current response. Interestingly, recent evidence suggests that rodents can perform similar tests that use sequential presentation of objects and tap into continuous recognition memory procedures (Kesner et al., 2001). In fact, 1-back and 2-back version of the human task have also been successfully used in rodents (Ko and Evenden, 2009). Note that continuous recognition paradigms offer another approach to test this capacity (Jackson-Smith et al., 1993; Kesner et al., 2001; Otto and Eichenbaum, 1992). Although it is clear that additional research is needed in this area, these findings support the idea that a more complete animal model of working memory will be available in the future.

2.4. Processing speed

Processing speed is defined in the NIH Toolbox CFB as “either the amount of time it takes to process a set amount of information or the amount of information that can be processed within a certain unit of time.” A pattern comparison task based on that designed by Salthouse is used to assess processing speed (Salthouse and Babcock, 1991; Salthouse, 1991). Pairs of stimuli are presented, and the subject must simply indicate, as quickly as possible, whether they are the same or different. To our knowledge, this task has not yet...
been developed for animals. Here, we discuss existing paradigms in primates and rodents that could be adapted to measure this subdomain.

2.4.1. Pattern comparison task

Pattern comparison tasks using morphed stimuli have been used to study feature-ambiguous discriminations in primates (Bussey et al., 2003; Saksida et al., 2006). The basic idea is to train animals to discriminate between 2 very different pictures, and subsequently measure performance as the 2 pictures are made progressively more similar (by morphing them). Although previous studies have focused exclusively on accuracy (Bussey et al., 2003; Saksida et al., 2006), the design could be updated to measure reaction time as well. It is important to note that age-related decline in speed of motor functioning may be a confounding factor in this test. However, this can be addressed by taking baseline measures of response time for each subject. A similar pattern comparison task could be developed in rodents using touchscreen testing methods (Bartko et al., 2011; Bussey et al., 2008).

2.4.2. Task validity and effects of aging

Distraction has a strong effect on tests of processing speed particularly in older adults, with irrelevant information influencing estimates of processing speed by more than 15% (Lustig et al., 2006). The task has been replicated to an extent in a canine model of aging in which performance accuracy declined with increased similarity of the distracter (Milgram et al., 2002). The study did not report response time and focused on errors alone, but the task described would be an excellent measure of speed of processing and would parallel the NIH Toolbox task.

2.5. Attention

The NIH Toolbox CFB recognizes the large array of processes that fall under the rubric of attention (e.g., selective [visual] attention, sustained attention, and divided attention) and focuses on sustained attention. However, attention also forms the basis for many conscious mental processes and overlaps with processes involved in executive function. The flanker task described in the NIH Toolbox CFB was designed to assess sustained attention but also response inhibition, a component of visual attention and executive function. It is based on a paradigm developed by Posner et al. (2002), and has been described previously.

We have examined 2 tasks that assess attention in animal models. The first task is a visual attention task, which has shown age-dependent performance effects in higher animal models. The second task is the 5-choice serial reaction test, which is extensively used in the literature to measure various aspects of attentional processes. Although the effects of attention on performance have been shown using this task, the other major selling point is the variety of cognitive measures that can be assessed using the single task. Both tasks are dependent on the frontal lobe and provide an excellent model to assess attention function in animals.

2.5.1. Five-choice serial reaction task

The 5-choice serial reaction (5-CSR) task is a rodent-based task designed to assess attention and several attentional processes. This task was originally developed (Carli et al., 1983) to assess attention-deficit/hyperactivity disorder (ADHD), and has been used to assess different aspects of attention in both humans and laboratory species. The test procedure consists of placing animals in the operant test box with a food delivery magazine on 1 side and a curved wall on the other. Five apertures are recessed in the curved wall on the same plane. The animal initiates a trial by a nose-poke into the magazine, which results in the onset of the intertrial interval (ITI; typically 5 seconds). Following the ITI, 1 of the 5 apertures is illuminated, and the animal is required simply to nose-poke in the lit aperture. A food reward is delivered for every correct nose poke. Errors result in a time-out phase (typically 5 seconds). The task typically lasts for 100 to 120 trials or 30 minutes, whichever is completed sooner. To manipulate the task load, it is also possible to use a “9-hole box.” This task provides a variety of response measures, including accuracy (proportion correct responses of total correct and incorrect responses, or percent correct), omissions (proportion of omitted trials, or percent omissions), anticipatory responses (premature responses, responses before stimulus presentation), perseverative responses (repeated responses at the response apertures), total trials completed, as well as several response latency indices. The 5-CSR is extensively used in the literature to assess attention deficits associated with neurodegenerative conditions (Harati et al., 2011; Muir et al., 1999; Romberg et al., 2011). Typically, selective attention is assessed by accuracy, whereas processing speed is assessed by measures of latency. Omissions measure sustained attention deficits (Young et al., 2009). This task has also been used to show that aging-induced deficits in attentional function is similar to that observed following lesions of the basalo-cortical cholinergic system (Muir et al., 1999). Furthermore, lesions of the hippocampus do not impair rat performance of the 5-CSRTT (Kirkby and Higgins, 1998), suggesting that there are similarities between brain regions that mediate human and rodent attention function.

2.5.2. Task validity and effects of aging

The forebrain cholinergic system shows a marked degeneration with age. Therefore, it should be expected that aged subjects show impairments in tests of attention. In 1992, Moore et al. (1992) attempted to assess the effects of aging in old Fisher-344 rats using simple- and choice-reaction time tasks (SRRT and CRTT, respectively), and reported age-related decrease in vigilance (Moore et al., 1992). Soon after, Jones et al. also demonstrated impairments in the 5-CSRT task using aged rats (Jones et al., 1995). Scopolamine has also been used to induce behavioral deficits in the 5-CSRT task, and AChEIs including tacrine, donepezil, and physostigmine all reverse scopolamine-induced deficits in performance (Kirkby et al., 1996). Recently, a computer-automated touchscreen version of the 5-Choice Serial Reaction Time Task (5-CSRTT) has been used to test the 3xTgAD mouse, which carries human APPsw and tauP301L transgenes as well as a PS1M146V knock-in (Romberg et al., 2011). These transgenic mice overexpress altered human genes, and many develop age-dependent learning and memory deficits. The authors report that the 3xTgAD mice were less accurate in responding to short, spatially unpredictable stimuli, particularly when the attentional demand of the task was high. Increase in perseverative responses was also reported. Another study has used the 5-CSRTT to report a prominent increase in premature responding (impulsivity) in transgenic mice expressing FTDP-17 tau mutations. These mice demonstrate filamentous neuropathology similar to that observed in human taulopathy conditions (Lambourne et al., 2007). The 5-CSRT can therefore be used to study behavioral specificity and to delineate specific behaviors relating to attention, executive function and speed of responding.

2.5.3. Visual attention task

Although the 5-CSRT is an excellent measure of attention function, it is possible to use another task that aligns more closely with the flanker task described in the NIH Toolbox CFB. However, it is important to note that perception of targets in humans is influenced a number of parameters, including image contrast and orientation of distractors in target detection tasks (Chen and Tyler, 2008; Polat and Sagi, 1993; Williams and Hess, 1998), so developing
an analog of the flanker task for other species is a significant challenge. The behavioral effects of flanking stimuli have mostly been studied in nonhuman primates (Li et al., 2006), although a recent study revealed that rats are also sensitive to the configuration of flanking elements of visual stimuli (Meier et al., 2011). More specifically, the authors report that rats are impaired in detecting collinear flankers, compared to configurations in which flankers differ from the target in orientation or angular position (Meier et al., 2011). Therefore, it would be an oversimplification to exactly imitate the flanker task in rodents, or in other laboratory species, if the correspondence in visual acuity were not fully characterized. However, a selective attention task can model the age-related attentional deficits by modifying the distracters used. Specifically the ability of young and old animals to respond to the target stimulus when it is presented along with 0, 1, 2, or 3 distracters is tested for the first sessions of the task. The difficulty of the task is further increased by modifying the distracter stimuli and using a distracter with greater physical similarity to the target for the later sessions of the task (Snigdha et al., 2012).

2.5.4. Task validity and effects of aging
In a recent study, we have shown that older dogs show decreased response accuracy coupled with increased response latency, and a greater disruption is seen when there is an increased similarity of distracter to the positive stimulus (Snigdha et al., 2012). This is consistent with studies in human subjects that reveal that challenging discrimination ability by using physically similar stimuli taxes attentional resources in older subjects (Scialfa et al., 1998) and in AD patients (Gainotti et al., 2001). These studies suggest that the visual attention task is a valid tool to predict age-related changes, as measured by decreased response accuracy coupled with increased response latency in animal models.

2.6. Language and reading
Language and reading involve the capacity that humans have for using a symbol system to categorize and communicate. In the NIH Toolbox CFB, there are 2 language measures, Oral Reading and Vocabulary Comprehension. Reading vocabulary, measured by oral pronunciation of single written words, is a surrogate for overall cognitive ability level but also for educational opportunity. Vocabulary Comprehension tests auditory comprehension of single words. Although some attempts have been made to measure communication in animals, this subdomain cannot be comprehensively investigated in animals and thus is not discussed in this article.

3. Conclusion
Animal experiments and preclinical testing have had an impact on our understanding of mechanisms of the aging process. However, their significance in predicting the effectiveness of treatment strategies in clinical trials has remained less than accurate. This is largely due to the translational gap between clinical and preclinical studies. This article attempts to draw attention to some of those gaps, in addition to identifying animal paradigms that closely parallel the tests included in the NIH Toolbox CFB. To parallel some of those gaps, in addition to identifying animal paradigms that challenge the behavioral effects of flanking sequences by twenty-month-old children. Child Dev. 64, 285–297.


