Refining the Cognitive Decathlon

Robert L. Simpson, Jr.
Applied Systems Intelligence, Inc.
3650 Brookside Parkway, Suite 500
Alpharetta, Georgia, USA

bsimpson@asinc.com

Charles R. Twardy
OnLine Star, Inc.
2515 Red Cedar Dr.
Bowie, Maryland, USA

ctwardy@onlinestarinc.com

ABSTRACT
We argue that cognitive tests of intelligent agents should use modern intelligence theory to help ensure the test battery covers key aspects of cognition and decomposes them as diagnostically as possible. To this end we assess the recent BICA cognitive decathlon proposal [15] on the Cattel-Horn-Carrol (CHC) factor model of human intelligence [11], and suggest tests to fill the gaps. Some of those tests come from cognitive performance software developed by NTI [17 & 18]. Appealing again to CHC theory, we note remaining gaps and suggest known tests which can fill them.

Categories and Subject Descriptors
I.2.0 [**Artificial Intelligence**]: General – *cognitive simulation*

General Terms
Measurement, Performance, Experimentation, Human Factors

Keywords
Cognitive decathlon, Integrated cognitive agents, Intelligence theory, Cattel-Horn-Carrol Model (CHC), BICA

1. INTRODUCTION
The idea of a cognitive decathlon dates back at least to the early 1990s when: "Vere proposed creating a "Cognitive Decathlon" to create a sociological environment in which work on integrated cognitive systems can prosper. Systems entering the Cognitive Decathlon are judged, perhaps figuratively, based on a cumulative score of their performance in each cognitive "event." The contestants do not have to beat all of the narrower systems in their one specialty event, but compete against other well-rounded cognitive systems." [23, p. 460]. In Newell [16] as well as Anderson and Lebiere [1], the goal is to resist specialization, and return AI to a broad vision of integrated intelligence. Anderson and Lebiere said their article could be viewed as a proposal for events in the decathlon, with initial scores provided by ACT-R and classical connectionism. Recognizing that goal, DARPA's Information Processing Technology Office (IPTO) has been looking for a "Cognitive Grand Challenge" to rival the highly-successful vehicle Grand Challenge. In January 2005, IPTO held a Grand Challenge workshop. They commissioned MITRE to prepare a report [2] detailing why previous Grand Challenge proposals had failed. Participants at the workshop were given copies of the report. It concluded that a Grand Challenge must meet these criteria:

- Clear and compelling demonstration of cognition
- Clear and simple measurement
- Decomposable and diagnostic
- Ambitious and visionary, but not unrealistic
- Compelling to the general public
- Motivating for researchers

These in turn were explained in some detail. For example, to be "clear and compelling":

a. The test should be a proxy for a range of problems requiring cognitive capabilities.
b. The test should not be "game-able" or solvable by "cheap tricks"
c. It should not be solvable by brute force computation, alone, and it should not lend itself to idiot savant solutions.
d. It should require integration of multiple cognitive capabilities.

The best general categories were "Physical Activity", like RoboCup, and "Take a Test". The MITRE review placed a Cognitive Decathlon into the "Take a Test" camp. But what sort of test?

2. RIGHT IDEA, WRONG TEST
RPI's Selmer Bringsjord [2, 4 &5] proposed that AI agents simply be given the Wechsler Adult Intelligence Scale (WAIS), a popular IQ test. Others proposed the New York Regent's exams, or the California STAR tests, which are performance tests, not aptitude tests. Bringsjord calls the general approach "Psychometric AI". Unlike the all-or-none Turing Test, failure on a *broad* test like WAIS *is* diagnostic – the pattern of successes and failures on the questions will tell us what the agent does well and poorly.

Furthermore, intelligence tests have been used for clinical diagnosis, opening up intriguing possibilities for "diagnosing" agents – which we expect will show a great many deficits when compared with humans, perhaps in characteristic patterns. For example, Paul Harrison [12] argues that statistical methods based on the Gaussian distribution react "autistically" to outliers.

However, the choice of test matters. WAIS and other standard tests are deficient because they cover mainly memory and attention, things which computers are very good at [9]. Indeed,

PerMIS'08, August 19–21, 2008, Gaithersburg, MD, USA.

Sanghi and Dowe [21] claim to have written a simple 960-line Perl program that gets average human scores on various IQ tests[1], which is clearly a *reductio* for those tests, since the program earns its scores on arithmetic, logical, and pattern questions, not language or semantic ones.

We argue that this is a problem with the *particular* tests, not the general idea. As we discuss in the next section, modern CHC theory holds that human intelligence factors into at least 10 broad aptitudes, only 2 or 3 of which are exercised by standard IQ tests. Flanagan et al. [10, p.54] claim, "*The Wechsler verbal/nonverbal model does not represent a theoretically or empirically supported model of the structure of intelligence*." Our proposal, in a nutshell, is to make sure that all are covered. We want the individual tasks to map fairly cleanly onto cognitive "modules" – cohesive units of cognitive function. So we need to know what those units are.

3. CHC: MODERN PSYCHOMETRIC THEORY

The underlying model of intelligence has changed in the hundred years since psychometric testing began.

Flanagan et al. [10 & 11] describe the progression of intelligence theories & tests from single-factor theories to modern theories. They argue that the most well-supported theory of cognitive factors is "modern G*f*-G*c*" theory. Their version is CHC theory, so called because it merges Carroll's 8-factor model based on an exhaustive review of the factor analysis literature[2] with the Horn-Cattell 10-factor model. CHC theory has 10 broad cognitive abilities, each of which subsumes between 2 and 14 more narrow abilities.

The most common intelligence tests (Stanford-Binet, and the Wechsler tests, including WAIS) *do not* match up with modern CHC theory. They were originally designed for single-factor or dichotomous theories of intelligence, and later revisions – the SB:IV and WISC-III or WAIS-III – have only been slightly updated: they do not correspond to the current consensus on the most likely cognitive factor/ability boundaries. Indeed, according to a new study, the recently-updated WISC-IV "measures [only] crystallized ability (G*c*), visual processing (G*v*), fluid reasoning (G*f*), short-term memory (G*sm*), and processing speed (G*s*); some abilities are well-measured, others are not" [13].

3.1 The CHC Factors

The factors are [17], pp.30-31, 42-45):

Gf – fluid intelligence: what we use when faced with a novel task; inductive and deductive reasoning.

Gc – crystallized intelligence: acquired knowledge; "the sage"

Gq – quantitative knowledge, esp. arithmetical

Grw – reading/writing ability; basic written comprehension & expression

Gsm – short-term memory; storage for a few seconds; working memory

Gv – visual processing: including spatial orientation

Ga – auditory processing: "the ability to perceive, analyze, and synthesize patterns among auditory stimuli, and discriminate subtle nuances in patterns of sound" (p.42)

Glr – long-term storage and retrieval: long-term memory performance (not content)

Gs – processing speed: "attentive speediness"; on the order of 2-3 minutes (p.44)

Gt – decision/reaction time or speed: on the order of seconds or parts thereof

To get a good measure of human cognitive abilities, CHC theory suggests at least two independent tests for each of these 10 broad abilities, preferably using relatively unrelated "narrow" abilities from within the broad ability. For example, a measure of Fluid Intelligence, G*f*, might include a test on "General Sequential Reasoning" and on "Induction".

Mueller et al. [15] present a cognitive decathlon they designed for DARPA's Biologically Inspired Cognitive Architecture program, BICA. BICA sought to "develop comprehensive biological embodied cognitive agents that could learn and be taught like a human." Mueller et al. developed three complex Challenge Scenarios, 23 Cognitive Decathlon[3] tasks, and a Biovalidity Assessment. We are concerned with the Decathlon tasks.

In Table 1, we have labeled the BICA decathlon tests according to the CHC abilities *we* think they measure. Unsurprisingly, the Visual tests measure Visual Processing, G*v*. Some of the more challenging ones may also measure long-term memory, G*lr*, given that they involve remembering and recognizing places previously visited. Likewise, the advanced Search tasks involve Processing Speed (G*s*), Memory (G*lr* & G*sm*), and possibly some Fluid Intelligence (G*f*) when the agent must learn hiding patterns.

Language and Knowledge areas test long-term memory (G*lr*) and Crystallized Intelligence (G*c*) and possibly Short-term Memory (G*sm*).

Part II of Flanagan et al. [10] is basically a how-to guide for constructing a minimal but sound cross-battery test to measure CHC abilities. Since we will see the cross-battery approach again with the NTI "Armory", we should consider Flanagan et al.'s core design ideas (pp.210-213)

• Use good theory (e.g. CHC) so you have good factors. That way we are more likely to cut at the joints, getting scores for each separate cognitive faculty, which is especially important in clinical settings, such as diagnosing learning difficulties.

• Use relatively pure indicators. Ideally, each task should measure a single factor, otherwise our indicator (the task score) contains reliable variance that is associated with another CHC construct, leading to confusion and misdiagnosis.

• Conversely, Use at least two distinct, qualitatively different narrow abilities to measure a broad ability. Otherwise you're not

[1] They did not attempt the Wechsler tests or the Stanford-Binet tests, presumably because they are not publicly available. Nor should we expect their program to do well on them. As we see later, those tests have heavy language and semantic components. Sanghi and Dowe used the ACE, Eysenck tests 1—8, I.Q. Test Labs, test Testedich.de, and an I.Q. test from Norway. They scored poorly on the last 3 (59, 84, and 60) respectively. See their Table 1 (p.4) and references.

[2] Carroll reviewed 1500 studies covering 461 data sets.

[3] No one interprets decathlon literally to mean 10 tasks.

measuring G*c*, but just VL or LD or LS, etc. (The Wechsler Verbal Comprehension Index is actually quite good in this regard.)

CHC theory describes a relative complete taxonomy of *cognitive* functions, but does not directly test abilities like attention, kinesthetic ability, causal understanding, tracking & timing, etc.

Table 1: Our Assessment of CHC Abilities Measured by the BICA Decathlon

Task	Level	CHC*
Vision	Invariant Object Identification	G*v*
	Object ID: Size discrimination	G*v*
	Object ID with rotation	G*v*
	Visual Action/Event Recognition	G*v*, G*lr*
Search & Navigation	Visual Search	G*v*
	Simple Navigation	G*v*, G*s*
	Travelling Salesman Problem	G*v*, G*s*, G*lr*
	Embodied Search	G*v*, G*s*, G*lr*
	Reinforcement Learning	G*v*, G*s*, G*lr*, G*f*, G*sm*
Manual Control & Learning	Motor Mimicry	--, G*sm*, G*v*
	Simple (1-hand) Manipulation	--, G*sm*, G*v*
	Two-hand manipulation	--, G*sm*, G*v*
	Device Mimicry	--, G*sm*, G*v*
	Intention Mimicry	--, G*sm*, G*v*
Knowledge Learning	Episodic Recognition Memory	G*lr*, G*sm*?
	Semantic Memory/Categorization	G*lr, Gf,* G*sm*?
Language & Concept Learning	Object-Noun Mapping	G*c*, G*lr*
	Property-Adjective	G*c*, G*lr*
	Relation-Preposition	G*c*, G*lr*
	Action-Verb	G*c*, G*lr*
	Relational Verb-Coordinated Action	G*c*, G*lr*
Simple Motor Control	Eye Movements	--
	Aimed manual Movements	--

* If presented verbally, all tasks also involve some auditory processing G*a*, and language, G*c*.

Manual Control and Simple Motor Control tasks test abilities outside the scope of CHC theory. However, some of the manual control tasks involve integrating a series of visual actions (part of G*v*) and remembering short sequences (G*sm*).

The BICA materials suggest that task instructions are presented verbally, in which case they also test auditory processing (G*a*) extensively. Nevertheless, if we are looking for more complete tests of cognitive ability in artificial agents, then CHC theory suggests we may want to supplement these decathlon entries with some that exercise other abilities:

G*f* – fluid intelligence

G*rw* – reading/writing ability

G*t* – decision/reaction time

G*s* – processing speed

G*q* – quantitative knowledge

Of these, perhaps the hardest to measure is G*f*. But we can make some progress with the others.

We might also be less interested in innate aptitudes than in cognitive *performance.* After all, cognitive systems are supposed to learn, so we might assess their capabilities at a specific time. For such repeated testing, it would be helpful to have a large "armory" of tests which can be composed on the fly. That armory idea comes from O'Donnell et al [17 & 18] at NTI.

4. THE NTI ARMORY

In this section, we look at a cognitive *performance* evaluation "armory" developed and computerized by NTI, Inc. of Fairborn, OH; see O'Donnell et al. [17 & 18]. The original goal of this effort was to permit researchers to generate unique test batteries from the armory that would be tailored to the performance demands of specific jobs for people. NTI reviewed existing taxonomies including the CHC, and created a list of 18 broad "performance attributes" or cognitive functions such as Sustained Attention, Working Memory, Decision Making, Spatial Visualization, and Time/Velocity estimation. The creation of an armory of tests that have been described in terms of a single defined set of performance and cognitive skills is noteworthy for our Cognitive Decathlon purposes. The NTI report summarizes a vast literature, and took a big step towards *applying* that literature to cognitive metrics.

Their idea was to rate each potential test/task against all of the 18 cognitive functions, creating a characteristic signature vector. The NTI software then creates an "optimal" test battery on the fly to match the skills needed by a particular job.

According to O'Donnell[4], "Since the armory was developed as a cognitive performance assessment tool, it has been used as a state measure, and has never been validated or compared to trait measures [such as CHC]. Some of the tests in the armory may have some history in the area of intelligence testing, but this was not our focus." However, we can use CHC categories even to guide performance assessments.

We envision a future Cognitive Decathlon web site where researchers could test the capabilities of their integrated cognitive agents by having their agents examined via administration of all or a subset of these and perhaps other tests. The advantage of this type of Decathlon is fairly straightforward. First, like CHC-based tests, these performance tests have a built-in comparison to human performance. Second, the tests are well understood within the

[4] Personal communication.

psychological testing community. Third, the average "man on the street" can understand the intuition of administering the same test to natural and artificial intelligent systems.

The NTI Armory is not sufficient for BICA's goals, particularly because someone could use a collection of subroutines each of which was optimized and specialized for a particular test, rather than an integrated cognitive agent. But our goal is to refine the BICA Decathlon, not replace it. CHC theory has led us to look for diagnostic tests that fill the gaps, and decompose relatively cleanly.

The specific tests in the NTI Armory are listed in Table 2. We provide a short description of some of these tests below.

Some tests – Dichotic Listening, Stroop Visual, and Visual Vigilance – depend on fine details of human cognition that we do not expect to see duplicated in machine cognition. For example, the famous Stroop test presents a color word like "red" in another color (as we did here, for those viewing this in color). The participant must try to name the color of the word, but humans find that difficult, and are prone to mistakenly say the conflicting color name from the word itself. (There is no difficulty with non-color words like "car".)

Table 2: The NTI Armory Tests

Continuous Memory	Reaction Time - Choice
Dichotic Listening	Reaction Time - Simple
Digit Span	Relative Motion (Join-Up)
Manikin (Low/High)	Sternberg - Letters
Match to Sample	Sternberg - Symbols
Math Processing	Stroop - Visual
Motion Inference	Tower of Hanoi (Low /High)
Novascan C (1 + 7)	Tracking - Pursuit
Precision Timing	Tracking - Unstable
Peripheral Information-Processing	Visual Vigilance
Rapid Decision Making	Wisconsin Card Sorting

In fact, a CMU-led team [8] showed that a simple neural net would generate human-like Stroop results so long as you trained it with more word-naming than color-naming tasks, so that word-naming was relatively automatic. That matched MacLeod and Dunbar's [14] showing that color naming itself was relatively automatic when paired with the even less well-trained task of shape naming, and that sufficient training on shape naming reversed that effect. So tests like the Stroop task are very good candidates for systems that learn like humans do.

Understanding the instructions may well be harder than taking the test itself. We do not want special-purpose agents that already know the task, so we must be able to describe the task to a general-purpose agent. The BICA proposal presumes a fair bit of verbal natural language processing (NLP). At minimum, agents would need to parse a formal language which can say, "You will get a task like this, and must remember x. Then you will get a distracter task where you have to do y, after which you will be asked to compare u and v to x."

Many of the NTI tasks put a lot of effort into directing human attention. As BICA imitation tasks like "do this" acknowledge, the ability to manage attention and indexical reference so the agent can have its attention directed is itself already a major achievement. The first round is likely to present tasks as the full set of percepts.

Let us now consider a few of the NTI tests. As our goal is not necessarily to exactly duplicate human performance, we should be prepared to use "staircase" techniques (e.g. [22]) to quickly find the system's limits, and then explore them. We recommend adding such features to almost all of the tests.

5. SOME NTI TESTS

5.1 Test 1: Continuous Memory

The continuous memory test consists of a random series of visual presentations of numbers which the operator must encode in a sequential fashion. As each number in the series is presented for encoding, a probe number is presented simultaneously. The operator must compare this probe number to a previously presented item at a pre-specified number of positions back in the series. Once the operator has made the appropriate recall, he or she must decide if that item is the same as, or different from, the probe number. Thus, the task exercises working memory functions by requiring operators to accurately maintain, update, and access a store of information on a continuous basis. Task difficulty is manipulated by varying the length of the series which must be maintained in memory in order to respond to recall probes.

Potential for Cog Decathlon: Obviously this is easy for a special-purpose program. However, it may still be a challenge for cognitive architectures like ACT-R which deliberately have a very limited working memory. For example, a recurrent neural net model will have a Markov horizon because computational constraints limit the chain depth. Also, any system operating "in the loop" with rich perceptions will be forced to limit attention and recall. We may have to bar some systems based on architecture, unless we can rely on a system gaming this test to fail another. Presenting the input "visually" (as images or feature vectors) is harder for "honest" systems, but still simple for special-purpose programs: just pipe the output of a trained digit classifier to a simple list processor, for example.

5.2 Test 4 and 5: Manikin

The Manikin Test as described here is a derivative of a task originally developed by [3] and popularized by the UTC-PAB [20]. The test is designed to index ability to mentally manipulate objects and determine orientation of a given stimulus. In this version, the test shows a vehicle such as an aircraft or a car. To one side is a male figure, and to the other side is a female figure. These figures and the object are lined up horizontally. Below the object and figures is a single query figure (male or female). The agent must determine whether the figure matching the query figure is to the right or the left of the vehicle, *in the vehicle's frame of reference* . The figures may appear either upright or upside down and facing either toward or away from the subject. The 16 combinations of orientation, stimuli and side are pseudo-randomly ordered. The number of trials selected for a given training or test session is under experimenter control. The NTI software uses stock images of the front or back of a sports car, and schematic man or woman figures (as you might see on restrooms, but in uniform). Unpracticed humans find most of the trials to be easy, but not when the vehicle is presented upside down and backwards. This test is considered to have two states in

which somewhat different cognitive skills are measured. In the LOW TRAINING condition, the subject is familiar with the task, but has not reached a level of "automaticity." In the HIGH TRAINING condition, the subject is so practiced that a different group of cognitive skills, such as procedural and working memory, are used to process the task.

Potential for Cog Decathlon: This requires visual presentation and an understanding of handedness. It is likely a test that machines would find difficult to do as fast as humans, since it involves a lot of image rotation, an understanding of "front/back/side", spatial awareness, and typical shapes of people and vehicles. It is still possible for a special-purpose program, but less so if the objects are chosen from a very large (and possibly unknown) set, and if we can apply obscurations to the image. Presuming a time limit, we can adapt this easily to test machines by reducing the time limit using a binary search. The metric for humans and machines could be the time-limit where they get 50% wrong.

5.3 Test 8: Motion Inference (Time/Velocity Estimation)

During the task, the subject sees a moving stimulus traversing a curved path. Approximately half way to a hash mark, the stimulus disappears. The subject's task is to determine when the stimulus, moving at a constant speed, would have reached a hash mark located in a random position along the curved path. The hash mark range of positions is set in the test's configuration program and can be anywhere between the beginning and end of the curve, but typically located in the last third of the path. The subject must infer how long the stimulus would take to reach the hash mark. The response required is a button press when the subject believes the stimulus would have reached the mark. The distracter is a simple "semantic" task. When the stimulus disappears, a series of four letters of the alphabet appear on the screen. The subject must immediately decide whether any of the letters are vowels. This decision is indicated with a response using a designated button on the response device (e.g., mouse). In effect, this interpolated task acts as a distracter to the subject in estimating the inferred motion. In this way, the subject is precluded from using methods such as counting, tapping, or singing to infer the motion. Once the response to the letters is made, the subject is required to estimate when the stimulus would have reached the stop point, and is to indicate this by pressing the designated button. This task really seems to require some practice.

Potential for Cog Decathlon: In addition to an interesting tracking task in itself, the distraction task forces us to consider how we present the directions to the cognitive agent. This is a good test, because it requires division and direction of attention between different tasks demanding different capacities. The distraction task could easily be gamed (if x in vowels: …), but once again, we seek other ways to prevent gaming. The visual tracking task should provide a challenge, and any agent capable of doing the visual tracking (given, say, a series of pixel planes) should be able to do visual inference of letters, which will make the task somewhat less trivial.

5.4 Test 9: NovaScan C

This test represents a special adaptation of the "multi-tasking" approach. Generally, in multi-tasking efforts the subject is free to adopt any strategy he or she wishes in order to achieve a final composite performance. This introduces some degree of difficulty in analyzing the task, particularly in diagnosing the nature of any decrements observed. NovaScan attempts to eliminate this ambiguity by using what has been called a "directed attention" rather than a "divided attention" paradigm. In the directed attention approach, the subject is still required to multiplex between two or more skill requirements. However, instead of being free to attend to each one whenever he/she wishes, the test directs the person to the test that must be attended to at any given time. This is done by having only one test appear on the screen at a time. In effect, the person has to keep one test's requirements in memory, while actively performing another test. In this way, the subject's strategy is highly constrained, and it is easier to determine where a cognitive decrement or improvement has occurred. Of course, it is still possible to introduce more than one task requirement at a time, as long as the demands of the tasks can be controlled.

NovaScan is a generic paradigm. There are many tests that can be introduced into it, just as there are many tests that can be used in the traditional divided attention approach. The present application of NovaScan, (C) uses two of the individual tests described elsewhere in the armory (Manikin and Continuous Memory). In each, a task appears on the screen (e.g., Manikin) and the subject must perform it for some period of time. At irregular intervals, this task is replaced by another task (e.g., Continuous Memory), and the subject must process this for some period of time. When that task is again replaced with the first (Manikin) task, the subject must remember the demands of the second task (Continuous Memory) while again performing the first. This alternation continues for some defined period of time or number of presentations. In addition to these demands, the subject typically must monitor a dial in which the pointer is moving at a constant rate, but in an inconsistent manner (the Dial Task). The subject must detect when the dial has gone into a "danger" zone. To do this, the subject must establish a scan rate for the dial that optimizes the opportunity to detect a danger indication, while allowing time to optimally process the other tests. This paradigm therefore approximates complex real-world tasks where two or more basic cognitive or psychomotor requirements must be attended to, and an optimal multiplexing strategy must be adopted based on current experience.

Potential for Cog Decathlon: The NovaScan paradigm offers a very flexible way to help prevent spoofing, since the agent must not only be able to do single tests, but switch between them. We should expect agents to have to learn the new combination, and then improve. Consider, for example, that learning to drive involves this kind of sequential directed attention, where subtasks are gradually automated. In fact, such considerations drove some of the early rule-generating systems. The instructions may still be the hardest part.

5.5 Test 12: Rapid Decision Making

The basic concept of this test is to present the subject with a display containing three "areas" that represent three levels of unspecified "danger". These areas are clearly marked with respect to the level of danger. At various times, symbols appear on the display indicating that a "vehicle" has entered into one of the areas. The vehicle appears as one of three types of symbol. One type clearly indicates that the vehicle poses minimal threat;

another indicates that the vehicle is a clear threat, and third type indicates that it is uncertain whether the vehicle is friend or foe. The subject's task is to decide on the level of threat, based on the type of vehicle and the area of the display in which it is located, and to make a differential response based on that decision. This is to be done as rapidly as possible. The test is paced so that only a short period of time is available to make the decision before the next stimulus appears, and this interval may be adjusted by the experimenter. In essence, this test is a complex choice reaction time test where higher level cognitive processes must be used to determine what the stimulus means, and where there is a complex response selection.

Potential for Cog Decathlon: This is obviously useful for a missile defense scenario. It naturally suggests a game where score is determined by a payoff matrix, where the values can be chosen at the start of the test. The main control variable is pacing. Other possibilities include number of locations and/or vehicle types. The uncertain vehicle is a nice touch, because optimal reward will require some utility calculations based on degree of certainty. The degree of uncertainty could be made to vary in a clear way, such as merging shape, or fading the image, or even just tagging it as uncertain and at what level.

5.6 Test 24: Wisconsin Card Sorting

In the armory's computerized version of this test, four groups of figures (called "key cards") are shown to the subject on the screen. Each card shows different shapes, and a different number of shapes. Also, the shapes on each card are a different color. They are typically arranged as shown in Figure 1, and this pattern of colors, shapes, and number is the default option.

Figure 1: Example shapes for the four "key" cards

The participant is then presented a series of "test cards" containing various combinations of the shapes, colors, and number of objects shown in the key cards. The task is to decide which key card "matches" the presented test card. Since there are three different ways a test card can match a key card (by color, shape, or number) the subject must decide which sorting criterion to use. No rule is given to the subject for matching cards. However, feedback is given for each attempted match on whether it was "right" or "wrong". This is based on a pre-established sorting criterion. Once the subject discovers the correct criterion and answers "correctly" six consecutive times, the criterion is switched to one of the other two. If the subject appears to be answering correctly for any number lower than six, and then makes an error, the count starts over (i.e., the subject must answer correctly six consecutive times). Normally, the types of shift in criterion are specified in the default condition. Among many dependent measures that may be collected, the number of matching categories completed and the number of "perseverative" errors (i.e., the number of matches attempted in which the same incorrect matching criterion was used) are perhaps most common. Perseverative errors indicate difficulty in changing approaches to problem solving, or inhibiting previously learned approaches. The task measures first the ability of the subject to conclude that there are 3 possible criteria by which to match the cards, and then assesses cognitive flexibility by requiring the subject to switch criteria to continue being successful at the task. The test is a good measure of adaptability and avoidance of perseveration.

Potential for Cog Decathlon: This card sorting task is a good test of a cognitive agent's ability to perform rule induction. It has the added twist that the rule has to be revised under some executive control when the rule is changed. An important control for comparison to human performance is the degree that the human has had prior experience with rule learning. This can be controlled somewhat with the prior knowledge that the cognitive agent has about the card representations. Like some of the other tests, e.g., the "join up" and "pursuit" tasks, if the agent has the appropriate learning mechanism this task should be easy and the advantage should be with the cognitive agent. Noise in the representation or other distracters could be added but then the difficulty goes up for the human subject perhaps beyond performance.

6. GAPS IN THE NTI ARMORY

The NTI Armory offers only partial coverage of all the potential dimensions of cognition. NTI's expert panel rated each test across all 18 of their defined cognitive functions. At least four dimensions of cognition are not well represented: Problem Sensitivity, Math Functioning, Language/Semantics, and Declarative Memory.

Problem sensitivity is the ability to recognize that a problem exists, not necessarily the ability to solve it. It is valued among emergency responders. Math functioning and language/semantics are self-explanatory. Declarative memory is memory of things from more than 20 minutes ago, hence a form of LTM but distinct from procedural LTM. So it would include both episodic (time-based) memory and other declarative (fact-based) memory.

We have identified several potential supplementary tests. To cover fact-based declarative memory, we could include the California Verbal Learning Test (CVLT). Also, using Table 5.1 from Flanagan et al. [10], the following items from common intelligence batteries are strong tests of LTM (G*lr*):

Tests of Associative Memory

- WJ-R Memory for Names
- WJ-R Delayed Recall Memory of Names
- WJ-R/III Visual-Auditory Learning
- WJ-R/III Delayed Recall Visual-Auditory Learning
- KAIT Rebus Learning
- KAIT Delayed Recall Rebus Learning

Tests of Ideational Fluency, Naming, or Declarative Memory

- WJ-III Retrieval Fluency (Ideational Fluency)
- WJ-III Rapid Picture Naming (Naming Facility)
- Visual paired-comparison (Declarative Memory)

Specific tests of language/semantics include the California Verbal Learning Test (CVLT) or any number of tests of crystallized intelligence from the WAIS and other intelligence batteries[5]:

- DAS Similarities

[5] List based on Flanagan et al. 2000a, Table 5.1.

- SB:IV Verbal relations
- SB:IV Comprehension
- SB:IV Absurdities
- WJ-III Verbal Comprehension
- WAIS Verbal Comprehension

To remedy the shortcomings in the Math Functioning dimension, we could include either the Wechsler-Arithmetic which contains 14 mental arithmetic brief story problems, the WJ-R/III Calculation and Applied Problems tests, or any number of similar tests of basic arithmetic. Story problems will require some language ability of course, while straight calculation tests could be trivial, if the agent can encode them directly into parseable code. We have been unable to identify a suitable test for problem sensitivity. An incident commander we spoke with suspects that this ability is usually assumed for emergency responders: training exercises will often require the responder to say, "Scene survey" and wait for the instructor to say "The scene is secure" or else fail, but no actual survey is performed.[6]

7. EPISODIC MEMORY

Episodic memory is declarative long-term memory (Glr) specifically associated with times or events – episodes – in an agent's history. It is a form of associative memory. Our ability to organize memories by events, such as yesterday's meeting or our last summer vacation, depends on (or exemplifies) episodic memory. The NTI test armory is weak here, especially because tests of long-term memory require, on their interpretation, 20-minute intervals. One of these tests would have to be paired with other tests that ran in the interval. However, the BICA tests are relatively strong. For example, one task requires the agent to remember which objects they have already encountered in which rooms.

There are some dedicated episodic memory tests. One is the University of Southern California Repeatable Episodic Memory Test [19]. It consists of:

...seven different lists, each co mposed of 15 semantically unrelated, high-frequency nouns. The words are presented in a different order on three study-test trials. After each study trial the subject recalls the words in any order. The test takes about 10 min to administer and score. The recall protocol can be scored for (a) global mnemonic efficiency, (b) primary and secondary memory, (c) subjective organizati on, (d) recall consistency and (e) recall as a function of serial position.

It has been applied in several clinical papers (for example, to Alzheimer's patients) to determine the pattern of memory deficiencies. Although it does not specifically require a 20-minute delay, it could. It is designed to be repeatable, and could be made even more so by using WordNet to generate lists of the required type on demand.

However, it requires a fair bit of semantic knowledge. Participants are expected to recall things by category, for example. Eventually, we want agents to be able to do this. In the meantime, however, we need a non-linguistic test of episodic memory. Several researchers in animal behavior (ethology) have been working on the problem.

[6] Bob Koester, personal communication.

Figure 2: Western Scrub-Jay caching or retrieving food. Scenery provides context for "episodic" memory. From N.S. Clayton, http://www.psychol.cam.ac.uk/cplcl/. Used with permission.

Researchers at Cambridge have been investigating food-cache storing in corvids, especially scrub-jays [6 & 7]. The design, as shown in the photo (Figure 2), involves a set of cache locations cued by features of the environment. The experimenters then compare the performance of caching birds, observing birds, and naïve birds on retrieval, attempting to control for various efficient search strategies.

A similar experiment could be set up as a software task (like those in the NTI armory), using successive still images or video. A simpler version could use very "cartoon" locations and stimuli. The agent being tested can then be asked where agent Green placed the items, or agent Blue. This would allow us to test episodic memory for agents that do not yet meet all of the BICA presumptions. (Of course, with cognitive agents it need not be food caching!)

8. CONCLUSION

In this paper we have reviewed the history of the idea of a cognitive decathlon as a methodology for testing the capabilities of an intelligent agent. We argued that the CHC criteria summarized by Flanagan et al.'s [10] presentation of modern intelligence theory – the Cattel-Horn-Carrol model lead nicely to specific cognitive categories for a decathlon. We also looked at a specific set of tests, the "NTI Armory," as candidates for a potential battery of tests. Admittedly, the NTI Armory offers only partial coverage of all the potential dimensions of cognition. We still need to complete the battery of tests for missing dimensions of cognition and describe how the tests would be administered to agent subjects.

9. ACKNOWLEDGMENTS

This paper is partially based upon work funded by DARPA and any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

10. REFERENCES

[1] Anderson & Lebiere 2003. The Newell Test for a theory of cognition. BBS 26:6, p.5.

[2] Bayer, S., Damianos, L., Hirschman, L. & Strong, G. A Summary of Previous Grand Challenge Proposals for Cognitive Systems. The MITRE Corporation. Prepared for DARPA IPTO, September 2004 Version 1.4 http://www.mitre.org/work/tech_papers/tech_papers_05/05_0947/index.html

[3] Benson, A. J., Gedye, J.L. and Jones, G. M. Disorientation in flight due to a covert vestibular disorder, with associated generalised, muscular tension. *Aerosp Med.* 1963 Jul;34:649–654.

[4] Bringsjord, S. & Schimanski. B. 2003. "What is Artificial Intelligence? Psychometric AI as an Answer," Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03) (San Francisco, CA: Morgan Kaufmann), pp. 887-893.

[5] Bringsjord, S. & B. Schimanski 2004. " 'Pulling It All Together' via Psychometric AI," Achieving Human-Level Intelligence Through Integrated Systems and Research Technical Report FS- 04-01 (Menlo Park, CA: AAAI Press), pp. 9-16.

[6] Clayton, N. S. and A. Dickinson. 1998. Episodic-like memory during cache recovery by scrub jays. *Nature* 295, 272—278.

[7] Clayton, N. S. and J. M. Dally and J. D. Gilbert and A. Dickinson. 2005. Food caching by Western scrub-jays (Aphelocoma Californica): a case of prospective cognition. *J. Exp. Psychol: Anim. Behav. Proc.* 31, 115—124.

[8] Cohen, J.D., Servan-Schreiber, D. and McClelland, J. 1992. A parallel distributed processing approach to automaticity. *American Journal of Psychology*, 105(2), 239—269.

[9] Cohen, P. R. 2005. If Not Turing's Test, Then What? AI Magazine 26(4): Winter 2005, 61–67

[10] Flanagan, D. P., McGrew, K. S. & Ortiz, S.O. 2000a. The Wechsler Intelligence Scales & Gf-Gc Theory: A Contemporary Approach to Interpretation. Allyn & Bacon, Boston.

[11] Flanagan, D. P., Ortiz, S.O., & McGrew, K. S. 2000b. Contemporary issues in intellectual assessment. Powerpoint presentation.

[12] Harrison, P. F. 2005. Is Autism a sensitivity to outliers? http://www.logarithmic.net/pfh/autism

[13] Keith, T., Fine, J., Taub, G., Reynolds, M. & Kranzler, J. 2006. Higher-Order, Multi-Sample, Confirmatory Factor Analysis of the Wechsler Intelligence Scale for Children—Fourth Edition: What Does it Measure? *School Psychology Review, 35 (1)*, 108-127. (From a review posted on McGrew's website, http://intelligencetesting.blogspot.com/2006/04/what-does-wisc-iv-measure-chc.html).

[14] MacLeod, C. M. & Dunbar, K. 1988. Training and Stroop-like interference: Evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 126—135.

[15] Mueller, M. S., Jones, M., Minnery, B. S. and Hiland, J. M. H. 2007. The BICA Cognitive Decathlon A Test Suite for Biologically-Inspired Cognitive Agents. BRIMS 2007, Norfolk, VA. http://obereed.net/docs/MuellerBRIMS2007.pdf

[16] Newell, A. (1990). Unified Theories of Cognition. Cambridge, MA: Harvard University Press.

[17] O'Donnell, R. D., Moise, S. and Schmidt, R. 2004. Comprehensive computerized cognitive assessment battery. Arlington, VA: Office of Naval Research; 2004. Contract No. N00014–01-C-0430. NTI, Inc.

[18] O'Donnell, R. D., Moise, S. and Schmidt, R. Generating performance test batteries relevant to specific operational tasks. *Aviation, Space, and Environmental Medicine* 2005; 76(7, Suppl.):C24–30. http://www.ingentaconnect.com/content/asma/asem/2005/00000076/A00107s1/art00007

[19] Parker, E. S., Eaton, E. M., Whipple, S.C., Heseltine and Bridge. 1995. *J Clin Exp Neuropsychol*. Dec;17(6):926-36. University of Southern California Repeatable Episodic Memory Test.

[20] Perez, W. A., Masline, P. J., Ramsey, E. G. and Urban, K. E. Unified Tri-Services Cognitive Performance Assessment Battery: Review and Methodology, Technical Report Apr 1984-Feb 1987. Systems Research Labs Inc., Dayton Ohio.

[21] Sanghi, P. and. Dowe, D. L. 2003. A Computer Program Capable of Passing I.Q. Tests, in P P Slezak (ed), Proceedings of the Joint International Conference on Cognitive Science, 4th ICCS International Conference on Cognitive Science & 7th ASCS Australasian Society for Cognitive Science (ICCS/ASCS-2003), 13-17 July 2003, Sydney, NSW, Australia, pp 570-575. http://www.csse.monash.edu.au/~sanghi/iq.html

[22] Watson, A. B. & Pelli, D. G. (1983) QUEST: a Bayesian adaptive psychometric method. *Percept Psychophys, 33 (2),* 113-20. http://www.psych.nyu.edu/pelli/pubs/watson1983quest.pdf

[23] Vere S. 1992. Planning, Encyclopedia of AI, Shapiro (ed.), Second Ed.