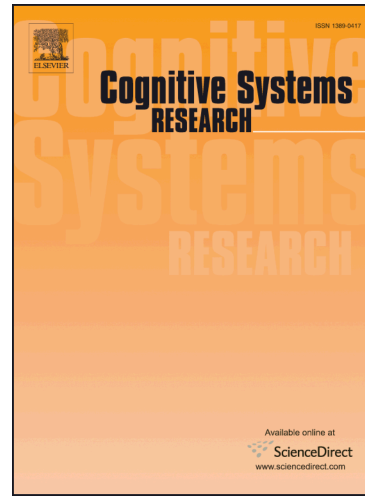

A Computational Analysis of General Intelligence Tests for Evaluating Cognitive Development

Fernando Martínez-Plumed, Cèsar Ferri, Jos'e Hernández-Orallo, María José Ramírez-Quintana

PII: S1389-0417(16)30129-2
DOI: http://dx.doi.org/10.1016/j.cogsys.2017.01.006
Reference: COGSYS 539

To appear in: *Cognitive Systems Research*

Received Date: 27 July 2016
Revised Date: 27 November 2016
Accepted Date: 23 January 2017

A Computational Analysis of General Intelligence Tests for Evaluating Cognitive Development

Fernando Martínez-Plumed, Cèsar Ferri, José Hernández-Orallo, María José Ramírez-Quintana
{fmartinez,cferri,jorallo,mramirez}@dsic.upv.es
Universitat Politècnica de València, Spain

Abstract

The progression in several cognitive tests for the same subjects at different ages provides valuable information about their cognitive development. One question that has caught recent interest is whether the same approach can be used to assess the cognitive development of artificial systems. In particular, can we assess whether the 'fluid' or 'crystallised' intelligence of an artificial cognitive system is changing during its cognitive development as a result of acquiring more concepts? In this paper, we address several IQ tests problems (odd-one-out problems, Raven's Progressive Matrices and Thurstone's letter series) with a *general* learning system that is not particularly designed on purpose to solve intelligence tests. The goal is to better understand the role of the basic cognitive operational constructs (such as identity, difference, order, counting, logic, etc.) that are needed to solve these intelligence test problems and serve as a proof-of-concept for evaluation in other developmental problems. From here, we gain some insights into the characteristics and usefulness of these tests and how careful we need to be when applying human test problems to assess the abilities and cognitive development of robots and other artificial cognitive systems.

Keywords: Cognitive development assessment; cognitive tests; general intelligence; task difficulty; cognitive operational constructs; general learning systems; evaluation of artificial systems, inductive programming.

1. Introduction

Humans undergo a cognitive development that starts with important neurological transformations even before birth and lasts during their lifetime. This cognitive development is also accompanied by an uneven variation of a range of cognitive capabilities. In order to assess this, many cognitive tests have been specifically devised for different capabilities and age ranges. In fact, one of the applications of these IQ tests (among many others) is the assessment of children's cognition and learning, in order to spot development problems or to identify specially talented individuals. The notion of 'mental age', for instance, was introduced by one of the fathers of psychometrics, Alfred Binet, to compare retarded children with the normal development of children of their age. Nowadays, the notion of a *single* mental age is more elaborate, as many abilities are known to develop at different age intervals.

This idea of using tests, in the form of a set of tasks or exercises, is also becoming more and more common for the evaluation of *artificial* cognitive systems or architectures. On one hand, we have those who advocate for the use of human cognitive tests for machines directly. This idea is behind the psychometric AI proposal (Bringsjord and Schimanski, 2003; Bringsjord, 2011), where sets of tests for all possible human capabilities should be used to evaluate artificial cognitive systems. This view is consistent with an editorial by Douglas K. Detterman (2011), the editor-in-chief of *Intelligence*, motivated by the success of IBM's program Watson (Ferrucci et al., 2010) on the *Jeopardy!* TV quiz show. Detterman challenged Watson and other artificial systems to pass a battery of human intelligence tests. For instance, *Spaun*, a 2.5-million-neuron model of the brain has been able to "reproduce the largest amount of functionality and behaviour" (Yong, November 2012) by their performance on a "diversity of tasks" (Eliasmith et al., 2012), where some of them are very similar to intelligence test tasks, such as serial working memory, counting, number series, etc. Despite this pullulation and advocacy of intelligence test tasks for evaluating artificial cognitive systems, some criticisms have been raised as whether the resulting scores can be meaningful (Dowe and Hernández-Orallo, 2012), starting with (Sanghi and Dowe, 2003a), who devised a very small and ad-hoc program that was able