UNIVERSIDADE FEDERAL DE MINAS GERAIS

HUDSON FERNANDES GOLINO

Validando Estágios de Desenvolvimento do Raciocínio Indutivo

Belo Horizonte
2012

HUDSON FERNANDES GOLINO

Validando Estágios de Desenvolvimento do Raciocínio Indutivo

Dissertação apresentada à Universidade Federal de Minas Gerais, como parte dos requisitos para obtenção do grau de Mestre em Psicologia.

Área de concentração: Desenvolvimento Humano.

Orientador: Prof. Dr. Cristiano Mauro Assis Gomes

Belo Horizonte
2012

150 Golino, Hudson Fernandes
G626v Validando estágios de desenvolvimento do raciocínio indutivo [manuscrito]
2012 / Hudson Fernandes Golino . -2012.

84 f. ; il.

Orientador: Cristiano Mauro Assis Gomes

Dissertação (mestrado) – Universidade Federal de Minas Gerais, Faculdade de Filosofia e Ciências.

1. Psicologia - Teses. 2. Psicologia do desenvolvimento – Teses. 3. Raciocínio – Psicologia – Teses. I. Gomes, Cristiano Mauro Assis. II. Universidade Federal de

AGRADECIMENTOS

Após muito trabalho, esforço e dedicação, me parece que a escrita dos agradecimentos é a parte mais dificil. Essa dificuldade se deve ao fato que, como qualquer conquista que obtemos na vida, nada se faz sozinho. Tenho que agradecer à muitas pessoas que, direta ou indiretamente, me ajudaram ao longo da minha trajetória desde o início da graduação até o mestrado. Deixo registrado o meu muito obrigado, de coração, à todos vocês que responderam ao teste, me ajudaram a corrigir e tabular, me incentivaram, instigaram, apoiaram ou que apenas estiveram presentes em algum momento importante.

Em especial, agradeço aos meus pais, Arnaldo Golino e Dinah Fernandes Golino. Obrigado por terem me ajudado a desenvolver um interesse tão grande pelo conhecimento e pela sua busca. Obrigado por me incentivarem e apoiarem em todos os momentos, desde o meu nascimento. Obrigado pelos conselhos, pelo absoluto bom humor, pela liberdade e confiança que sempre tiveram, pelos puxões de orelha tão merecidos, por todo o esforço que tiveram para sustentar e financiar os meus estudos, e por todo o *bullying* familiar que me ajudou a calejar o caráter. Agradeço, também, ao meu irmão Diogo por se alegrar, talvez mais que eu, com cada etapa vencida ou realização atingida, e por toda a ajuda que sempre me deu.

Minha carreira não existiria sem a presença do Prof. Cristiano. Acho muito difícil agradecer em poucas linhas as incontáveis horas de orientação e supervisão, sempre tão ricas e significativas, assim como todo o esforço para me ajudar a desenvolver um raciocínio preciso, profundo, crítico e autônomo que me leva cada vez mais adiante, com a liberdade que eu tanto prezo. Muitíssimo obrigado por toda a jornada desde 2008, e pelas próximas que perseguiremos. Não tenho como deixar de agradecer a paciência e compreensão da Flávia Schayer Dias e da Isabella Schayer Dias Assis, pelas horas dispendidas em sua casa discutindo idéias ou fazendo análise de dados morosas, algumas vezes em horário de descanso.

Agradeço aos colegas de pós-graduação, Bianca Costa, Thiago Vasconcelos, Hunayara Tavares, pela companhia e troca rica de idéias durante este percurso, e à Michele Ferreira e Cristiane Gomes pela grande ajuda com a coleta e tabulação de dados do ano de 2012! Agradeço também à Sheila Couto, Marília Souza e Eunaihara Marques pela companhia nos últimos anos.

Tenho de registrar aqui um agradecimento especial aos meus amigos de longa data, peças importantes na minha vida, em especial ao Igor Thomaz, Gabriel Bernardes, Pedro Pires, Alberto Mello, Alexandre Braga e Bernardo Melo. A presença de vocês em cada momento de todos esses anos foi de fundamental importância. A junção de pessoas fora do padrão sempre contribui para crermos, erroneamente, que somos normais.

Agradeço ao Ricardo Aguiar e Uriatan Benevides Dutra, pela presença marcante e contínua, e por todo o auxílio em meus treinamentos. Agradeço à todos os companheiros e amigos da Equipe Shigueto, em especial o Professor Cláudio Rodrigues, por terem me ajudado a manter uma excelente qualidade de vida e controle de estresse no último ano, por meio da arte suave.

Agradeço aos Professores Igor Gomes Menezes e Ângela Maria Vieira Pinheiro por terem aceito compor as bancas de qualificação e de defesa do Mestrado, apesar da agenda corrida que têm. Obrigado pelas sugestões, correções, apontamentos, orientações e pela parceria profícua. Agradeço-os imensamente pela paciência com os trâmites burocráticos, e com as marcações sempre de última hora, sem a flexibilidade que tiveram seria difícil fazer a defesa no prazo que escolhi, sete meses após o início do mestrado.

Deixo registrado o meu agradecimento à Fapemig pelo financiamento de parte desse projeto, assim como à Fundep por ter tornado a administração do recurso algo simples e objetivo para mim e para o Prof. Cristiano. Agradeço, também, a agilidade profissional e disponibilidade da Flávia e do Fabrício, profissionais que tornam a vida de todos os alunos do Programa de Pós-Graduação em Psicologia da UFMG mais tranquila.

Agradeço aos amigos e colaboradores Patrice Marie Miller e Michael Lamport Commons, por toda ajuda e oportunidades que têm me oferecido nos últimos anos.

À minha noiva Mariana Teles Santos, por ter chegado no momento certo...

Epígrafe

"A page sheet of music *represents* a piece of music; the music itself is what you get when the notes on the page are sung or performed on a musical instrument. It is in its performance that the music comes alive and becomes part of our experience; the music exists not on the printed page but in our minds. The same is true for mathematics; the symbols on a page are just a *representation* of the mathematics. When read by a competent performer, the symbols on the printed page come alive – the mathematics lives and breathes in the mind of the reader like some abstract symphony. Furthermore, mathematics makes the invisible visible."

Keith Devlin

RESUMO

A validade dos estágios de desenvolvimento é uma questão controversa na literatura sobre Psicologia do Desenvolvimento. No entanto, nos últimos vinte anos, uma série de metodologias quantitativas têm sido empregadas para se verificar, empíricamente, a existência de descontinuidades, tanto no desempenho das pessoas quanto nas dificuldades de itens e tarefas. A dissertação tem como foco a investigação acerca da validade de estágios de desenvolvimento do raciocínio indutivo, por meio da construção e validação do Teste de Desenvolvimento do Raciocínio Indutivo (TDRI). Ela está dividida em duas partes, que representam dois artigos. O primeiro apresenta as duas versões iniciais do TDRI, e investiga se os itens mensuram os estágios de desenvolvimento, formando grupamentos distintos entre si, em duas amostras, uma composta por 167 pessoas (50.3% homens) com idades entre 6 e 58 anos (M = 18,90, DP = 9,70), e a outra composta por 188 pessoas (57.7% mulheres) com idades entre 6 e 65 anos (M = 21,45, DP = 14,31). Os resultados apontam um adequado ajuste ao modelo dicotômico de Rasch (infit médio = 0,94; desvio-padrão = 0,22), e evidenciam que os grupamentos de itens seguem o padrão previsto (oito itens por grupamento, cada grupamento formando um estágio), e que grupamentos adjacentes apresentam diferenças significativas entre si. O segundo artigo investiga a validade estrutural da 3ª versão do TDRI, que foi construída para superar algumas limitações verificadas nas primeiras duas versões. Esse segundo estudo emprega três metodologias distintas para verificar a validade dos estágios de desenvolvimento: 1) Análise Fatorial Confirmatória (AFC); 2) Análise Rasch para dados dicotômicos; e 3) Análise de classes latentes. A amostra foi composta por 1.459 pessoas people (52.5% mulheres) com idade entre 5 e 86 anos (M = 15,75, DP = 12,21). O resultado aponta uma estrutura fatorial de dois níveis, sendo o primeiro nível composto por 7 fatores (um para cada estágio) e o segundo nível um fator geral [$\chi2$ (61) = 8832.594, p = .000, CFI = .96, RMSEA = .059]. Os itens do TDRI se ajustam ao modelo Rasch (infit médio = 0,96; desvio-padrão = 0,17), e apresentam uma confiabilidade alta para os itens (1.00) e moderadamente alta para as pessoas (0,82). As evidências apontam que a solução com sete classes latentes apresenta o melhor ajuste aos dados (AIC: 263.380; BIC: 303.887; Loglik: -111.690). Os estudos que compõe essa dissertação mostram que é possível, a partir da adoção de uma série de metodologias específicas, identificar empiricamente estágios de desenvolvimento.

As evidências apontam que o TDRI é um instrumento válido e confiável para avaliar estágios de desenvolvimento do raciocínio indutivo.

Palavras-chaves: validade estrutural, estágios, desenvolvimento, raciocínio indutivo.

ABSTRACT

The developmental stages validity has been focus of controversy in the literature about developmental psychology. However, in the past twenty years a serie of quantitative methodologies have been developed or applied to empirically identify discontinuities, both on people performance and items and tasks' difficulties. The present dissertation investigates the validity of inductive reasoning developmental stages, throught the construction and validation of the Inductive Reasoning Developmental Test (IRDT). It will be presented in two parts, representing two papers. The first paper investigates if the IRDT's items measures six developmental stages, forming six different and spaced clusters, in two samples, being one composed by167 people (50.3% men) with ages varying from 6 to 58 years (M = 18.90, SD = 9.70), and the other composed by 188 people (57.7% woman) with ages varying from 6 to 65 years (M = 21.45, SD = 14.31). The result shows an adequate data fit to the Rasch model (infit mean = 0.94, SD = 0.22), six clear item clusters with gaps between them, with adjacent clusters presenting statistically significant differences. The second paper investigates the structural validity of the IRDT 3rd version, constructed to overcome some limitations founded in the first two versions. Three quantitative methodologies are used: 1) Confirmatory Factor Analysis; 2) Dichomotomous Rasch Model; and 3) Latent Class Analysis. The sample was composed by 1,459 people (52.5% woman) aging from 5 to 86 anos (M = 15.75, SD = 12.21). The result shows a factorial structure with seven first-order latent variables (one for each stage) and a second-order geral factor [$\chi 2$ (61) = 8832.594, p = .000, CFI = .96, RMSEA = .059]. The 56 items presented adequate fit to the Rasch model (infit mean = 0.96; SD = 0.17), with a high item reliability (1.00) and a moderately high person reliability (0.82). The evidences point to a seven latent class model (AIC: 263.380; BIC: 303.887; Loglik: -111.690). Both studies show that is possible to empirically identify developmental stages of reasoning applying specific quantitative methodologies. The evidences point to the validity of the IRDT items to assess developmental stages of inductive reasoning.

Keywords: structural validity, stages, development, inductive reasoning.

LISTA DE TABELAS

Tabelas do artigo 1

Tabelas do artigo 2

LISTA DE FIGURAS

Figuras do artigo 1

Figuras do artigo 2

LISTA DE ABREVIATURAS E SIGLAS

IRDT	Inductive Reasoning Developmental Test
Infit	Information weighted fit statistics
DST	Dynamical Skill Theory
MHC	Model of Hierarchical Complexity
HCSS	Hierarchical Complexity Score System
LAS	Lectical Assessment System
OHC	Order of Hierarchical Complexity
SLM	Simple Logistic Model
Pre-op/SR	Pre-operational or Single Representations
Prim/RM	Primary or Representational Mappings
Conc/RS	Concrete or Representational Systems
Abst/SA	Abstract or Single Abstractions
Form/AM	Formal or Abstract Mappings
Syst/AS	Systematic or Abstract Systems
Abs	Abstract Items
Sys	Systematic items
Met	Metassystematic items
CFA	Confirmatory Factor Analysis
LCM	Latent Class Models
AIC	Aikaike's information criterion
BIC	Bayesian information criterion
EM	Expectation-maximization algorithm
CFI	Comparative Fit Index
RMSEA	Root Mean Square Error of Approximation

SUMÁRIO

1. APRESENTAÇÃO

A idéia de que o ser humano se desenvolve por meio de estágios específicos é alvo de grande controvérsia e debate na literatura em Psicologia do Desenvolvimento (Miller, 2002; Morra, Gobbo, Marini, & Sheese, 2008). Apesar de ter sido muito influente em grande parte do século XX, a partir de 1980 a noção de estágios começou a entrar em declínio, devido à dois fatores principais: 1) Um corpo crescente de evidências que levaram alguns pesquisadores a afirmar que era uma teoria inapropriada de desenvolvimento cognitivo (Morra et. al, 2008), e 2) Críticas abordando questões filosóficas e epistemológicas acerca da noção de estágios (Marshal, 2009).

Apesar do debate e das controvérsias que ainda se encontra na literatura especializada, Fischer e seus colaboradores mostram que a identificação tanto de desenvolvimento descontínuo quanto de desenvolvimento contínuo é uma questão de foco de análise e de metodologia (Fischer, Kenny, & Pipp, 1990; Fischer & Silvern, 1985; Fischer & Yan, 2002a,b; Schwartz & Fischer, 2005; Yan & Fischer, 2007). O desenvolvimento contínuo diz respeito à sequência de passos ou procedimentos necessaries na construção das habilidades (microdesenvolvimento), enquanto a descontinuidade diz respeito à mudanças abruptas, do tipo estágio, que demarcam a emergência de novos tipos de controle de unidades do comportamento e da cognição (Fischer, 1980; Fischer & Rose, 1994; Fischer & Bidell, 1998, 2006; Fischer & Yan, 2002a).

A identificação empírica de estágios de desenvolvimento faz parte da agenda de pesquisas de um grupo de pesquisadores pós-piagetianos, que nos últimos trinta anos tem desenvolvido ou aplicado novas metodologias e técnicas que tornam possível a verificação de descontinuidades, tanto em termos de habilidade das pessoas quanto em dificuldade dos itens. Evidências robustas de estágios de desenvolvimento têm sido apresentadas por meio da aplicação dos modelos Rasch, analisando-se a distribuição dos itens ao longo da escala da variável latente (Dawson, 2000; Dawson, Xie, & Wilson, 2003; Bond & Fox, 2001; Müller, Sokol, & Overton, 1999), verificando-se as curvas características dos itens (Dawson-Tunik, 2004; Dawson-Tunik, Commons, Wilson & Fischer, 2005), utilizando-se estatísticas univariadas, como o teste t de Student, para verificar diferenças entre grupamentos de itens (Bond & Fox, 2001; Commons et al., 2008; Dawson, 2002; Golino, Gomes, Commons, & Miller, in press) e por meio da utilização de análises de classes latentes (Bond & Fox, 2001; Dawson-Tunik et. al., 2010; Demetriou & Kyriakides, 2006).

Além da utilização de metodologias estatísticas sofisticadas para identificar estágios de desenvolvimento, o emprego de modelos matemáticos de organização de informação também se constitui como um caminho que tem se mostrado eficaz para a mensuração de estágios, uma vez que se constituem como um guia de referência para a construção de itens e tarefas (Commons et al, 2008; Dawson-Tunik, 2004; Dawson-Tunik, Commons, Wilson, & Fischer, 2005). O modelo matemático de organização de informações mais bem formulado e claro é o Modelo da Complexidade Hierárquica – MCH – (Commons, 2008; Commons & Pekker, 2008; Commons, Gane-McCalla, Barker, & Li, no prelo), que se insere na chamada Teoria Matemática da Medida (Krantz, Luce, Suppes, & Tversky, 1971; Luce, & Tukey, 1964).

A presente dissertação tem como objetivo o estudo de validade de estágios de desenvolvimento do raciocínio indutivo por meio da junção do MCH, que serviu como referência para a construção do Teste de Desenvolvimento do Raciocínio Indutivo, com as metodologias quantitativas utilizadas para se verificar estágios de desenvolvimento. Dois estudos foram conduzidos. No primeiro estudo, apresentamos as duas versões iniciais do teste, e utilizamos o modelo logístico simples de Georg Rasch para verificar se os itens seguiam o padrão de dificuldade predito pelo MCH. Nesse estudo, as evidências de estágio são investigadas por meio do grupamento de itens com mesmo grau de complexidade, verificando se diferenças entre grupamentos (estágios) adjacentes são estatísticamente significativas. No segundo estudo, aprimoramos o teste e utilizamos três metodologias quantitativas diferentes, cada uma buscando investigar um aspecto específico da validade estrutural do teste. A análise fatorial confirmatória busca explorar a estrutura (ou arquitetura) dimensional do instrumento, que é esperado apresentar sete fatores de primeiro nível e um fator geral de segundo nível. A análise Rasch busca verificar se itens construídos para mensurar um estágio específico se agrupam, e se grupamentos diferentes de itens estão espaçados ao longo do traço latente. O test t de Student é utilizado para verificar se esses espaçamentos são significativos. Por último, utilizamos um modelo de classe latente, a fim de verificar quantas variáveis latentes discretas explicam a distribuição de dificuldade dos itens. Cada metodologia proporciona informações diferentes e complementares sobre os estágios, e juntas podem formar um conjunto de evidências mais robusta do que a utilização de uma ou outra metodologia isoladamente.

O primeiro artigo foi aceito para publicação em uma edição especial da *Journal of Adult Development*, a ser lançada no ano de 2013. No entanto, uma publicação online prévia já pode ser encontrada no sequinte link: http://adultdevelopment.org/jad_special_issue.php. O segundo artigo será submetido ao *International Journal of Testing*. Cada artigo possui formatação específica, de acordo com a revista alvo. Como preferimos montar a dissertação em forma de dois artigos, o leitor irá se deparar com duas formatações diferentes.

2. ARTIGOS

2.1 Artigo 1

The Construction and Validation of a Developmental Test for Stage Identification: Two Exploratory Studies

Part of this research was supported by the *Instituto Ester Assumpção*, and by the *Fundação de Amparo à Pesquisa do Estado de Minas Gerais* (FAPEMIG). We are thankful to all those involved in the revision of the manuscript, in special Prof. Cory David Barker (Antioch University Midwest), Prof. Igor Gomes Menezes (UFBA, Brazil) and Prof. Ângela Maria Vieira Pinheiro (UFMG, Brazil), for all the suggestions.

Abstract

The present work presents two exploratory studies about the construction and validation of the Inductive Reasoning Developmental Stage (IRDT), a forty-eight items test based on the Model of Hierarchical Complexity. The first version of the test was administered to a convenience sample composed by 167 Brazilian people (50.3% men) aged between 6 to 58 years (M = 18.90, SD = 9.70). The Rasch Model was applied, and the result shows reliability of .97 for the full scale. The Infit mean was .87 (SD = .28; Max = 1.69; Min = .39), and the person reliability was .95. One sample t-tests showed that the spacing of Rasch scores between items of adjacent orders of hierarchical complexity is significant, with large effect size. The second study was conducted in order to overcome some of the test's limitations found in the first study. The revised IRDT were administered to a convenience sample composed of 188 Brazilian people (57.7% women) aged between 6 to 65 years (M = 21.45, SD = 14.31). The reliability for the full scale was .99, and its Infit mean was .94 (SD = .22; Max = 1.46; Min = .56). The person reliability was .95. One sample t-tests showed that the spacing of Rasch scores between items of adjacent orders of hierarchical complexity is significant, with large effect size. The paper finishes with a discussion about the necessity and importance to focus on the vertical complexity of the items in any test designed to identify developmental stages.

Keywords: Stages, Assessment, Validation, Development, Model of Hierarchical Complexity, Inductive Reasoning.

The Construction and Validation of a Developmental Test for Stage Identification: Two Exploratory Studies

Piaget is considered one of the most important researchers of the 20th century (Flavell, 1963), with his studies creating a very influential framework within developmental psychology, that of Genetic Epistemology. In spite of its importance, the influence of the theory on developmental research began to decline in the 1980's, due to a large body of evidence that apparently contradicted the theory's notion of developmental stages (Marshal, 2009; Miller, 2002). One might say that this theory was "put in check" by the maneuvers of others. When Piaget's theory, specifically his stage concept, was put in check, all Piagetian and Neo-Piagetian developmentalists were, in some manner, placed in the same condition. As in chess, getting out of the check is of great importance, and requires the development and implementation of sturdy strategies. In developmental psychology, getting out of check can be reached through the implementation of "strategic moves", as in the construction of better metrics (Fischer & Rose, 1999; Rose & Fischer, 1998; Van Geert & Steenbeek, 2005), with reliable, valid and accurate measures (Fischer & Dawson, 2002), and the adoption of quality control standards (Stein & Heikkinen, 2009).

The current paper presents one of these moves which, together with other work~~s~~ (Commons, Trudeau, Stein, Richards, & Krause, 1998; Commons et al., 2008; Dawson, 2003, 2006; Dawson & Wilson, 2004; Dawson, Goodheart, Wilson, & Commons, 2010; Dawson-Tunik, Commons, Wilson, & Fischer, 2005; Demetriou & Kyriakides, 2006; Fischer, 2008; Fischer & Bidell, 1998, 2006; Rijmen, De Boeck, & Van der Mass, 2005; Van der Maas & Molenaar, 1992), aims to collaborate in getting out of the check. Two exploratory studies about the construction, challenges and initial results of the Inductive Reasoning Developmental Test (IRDT) - *Teste de Desenvolvimento do Raciocínio Indutivo* (Gomes & Golino, 2009) will be presented. The IRDT intends to measure developmental stages of inductive reasoning through reliable, valid and accurate measures, falling in the category of so-called "quality control standards".

Criticisms of Stages, or Killing Piagetian Stage Theory:

Beginning in the 1980's, increasing numbers of researchers began to criticize Piagetian stage theory (Miller, 2002; Morra, Gobbo, Marini, & Sheese, 2008). The main criticisms were directed at the idea that stages are structures of the whole, developing in a synchronous way, emerging at specific ages, and reaching a single *telos*, represented by formal operations (Fischer & Bidell, 2006).

One set of criticisms that emerged empirically supported the idea that variability is the norm, rather than the exception in human development (Bidell & Fischer, 1992, 2006; Fischer & Rose, 1999; Flavell, 1963; Miller, 2002; Siegler, 1981). Such evidence~~s~~ points to asynchrony, heterogeneity and high variability in performance (Demetriou, Efklides, Papadaki, Papantoniou, & Economou, 1993; Fischer & Bidell, 2006). Some major studies indicate *decaláge* in the ability of seriation (Chapman & Lindenberger, 1988; Halford, 1989; Jamison, 1977), conservation (Kreitler & Kreitler, 1989; Nummedal, 1971; Murray, 1969; Murray & Holm, 1982), formal operations (Bart, 1971; Lautrey, de Ribaupierre & Rieben, 1985; Martorano, 1977; Webb, 1974), combinatorial analysis (Roberge, 1976; Scardamalia, 1977), object permanence (Baillargeon, 1987; Chazan, 1972; Jackson, Campos & Fischer, 1978), among others.

In addition to studies showing massive *decaláges*, age issues and synchronism problems on Piagetian theory of cognitive development, other revisions of the theory were made. Commons and Richards (1984a), Commons, Richards and Kuhn (1982), Fischer (1980, 1987), Fischer, Hand and Russell (1984), and others, argued that the stage of formal operations is not the last possible level in human cognitive development, and show evidence for post-formal levels.

The other set of criticism emerged from philosophical/epistemological positions. Broughton (1984), for example, argued that formal operations are a wholly inadequate model of thought in adolescence and adulthood, and as a result believes the entire theory should be reconsidered.

The criticism, sometimes based on empirical aspects, sometimes based on philosophical and epistemological positions, was striking, and came from many different lines. Flavell even in his early work entitled *The Developmental Psychology of Jean Piaget* (1963), points to ambiguities in the concept of stage, argues about the challenges of the clinical method, on the impossibility of stating that a child "has" a particular concept and raises the question of language as an intervening variable (Siegler & Crowley, 1991). Despite recognizing the historical importance of Piaget’s work, in particular the stage theory, he comes to argue, in another, later work, that the Piagetian stage theory “explains nothing" (Flavell, 1985; Lourenco, 1998). Lourenço (1998) proposed that many cognitivists (e.g. Bjorklund, 1997; Brainerd, 1997; Cohen, 1983) already considered Piaget’s theory to be dead, and some of them suggested that there was no real purpose in continuing to test a theory that was already known to be inadequate (Halford, 1989; Lourenco, 1998).

In short, until the mid 80’s the classic structuralism of Piaget’s theory had significantly influenced developmental psychology research worldwide (Marshall, 2009). In spite of being one of the most important players of the “Developmental Chess,” the grandmaster was double *checked*. His influence, including the concept of stages, began to decline, due mainly to (1) the growing body of evidence that helped convince some researchers that stage theory was inappropriate to describe cognitive development (Morra, et al., 2008), and to (2) criticisms that addressed philosophical issues and suggested an epistemological reconfiguration (Marshal, 2009).

Neo-Piagetians and Post-Piagetians

A group of Neo-piagetian researchers has sought to overcome the problems and limitations pointed to in the Piagetian concept of stage, including his methodology for assessing them, proposing instead more modern theoretical and methodological approaches that have been providing new evidence~~s~~ for discontinuity. Included in these newer approaches are two important and related models of development: Fischer’s Dynamic Skill Theory (DST) and Commons’ Model of Hierarchical Complexity (MHC). Fischer (1980) proposed a set of analytical tools that make possible the detailed description of developmental pathways, as well as the construction of domain-free hierarchical taxonomies to classify performance. His DST (Fischer, 1980; 2008; Fischer & Bidell, 1998, 2006; Fischer & Rose, 1994, 1999; Fischer & Yan, 2002a,b) conceives of development as a phenomenon composed of both continuous and discontinuous patterns of changes. The former (continuous change) relates to the sequence of steps followed in the construction of skills (microdevelopment), while the latter (discontinuous change) relates to abrupt, stage-like changes that marks the emergence of radically new kinds of control units of behavior and cognition (Fischer, 1980; Fischer & Rose, 1994; Fischer & Bidell, 1998, 2006; Fischer & Yan, 2002a). Evidence for both kinds of developmental patterns have been shown by Fischer and colleagues (Fischer, Kenny, & Pipp, 1990; Fischer & Silvern, 1985; Fischer & Yan, 2002a,b; Schwartz & Fischer, 2005; Yan & Fischer, 2007). Instead of conceptualizing the discontinuous facet of human development as a unidirectional ladder, however, the DST conceptualizes it as a *constructive web* that encompasses the activity of the person and the supportive context in which this action is performed (Bidell & Fischer, 1992; Fischer & Bidell, 2006). So, a person may have a certain level of performance, let us say X, in the domain of Algebra, and an X-1 level of performance in the domain of Combinatorial Analysis, for example. Furthermore, this same person may present higher or lower levels of performance in the previously cited domains due to social support (scaffolding), emotional reactions, and so on (Fischer & Bidell, 2006). The *constructive web* notion is different from the Piagetian concept of stages as developmental ladder, in which *decalage* is the exception.

Despite the importance and contribution of the DST to the Developmental Sciences field (Miller, 2002; Morra et. al, 2008), it was Commons and his colleagues that have proposed the groundwork for the mathematical formalization of discontinuity, through the Model of Hierarchical Complexity (MHC). The MHC is a general measurement theory, and as such is part of the normal Mathematical Theory of Measurement (Krantz, Luce, Suppes, & Tversky, 1971; Luce, & Tukey, 1964) applied to the phenomenon of difficulty. The MHC introduces the concept of the Order of Hierarchical Complexity (OHC) that conceptualizes information in terms of "the power required to complete a task or solve a problem" (Commons, Trudeau, Stein, Richards, & Krause, 1998). Commons and Pekker (2008) demonstrated, in axiomatic terms, that task difficulty or complexity, beyond other sources, increases in two ways: horizontally and vertically. The first refers to the accumulation of informational bits necessary to successfully complete a task (Commons, 2008), e.g. 5 + 6 + 7 is less complex than 5 + 6 + 7 + 8, because the first differs from the second in the number of times addition was executed, and does not differ in the organization of the addition itself; that is, both have the same *hierarchical (or vertical) complexity.* So, horizontal or traditional complexity is just the adding of informational bits. Vertical complexity, or *hierarchical complexity*, refers to the organization of information in the form of action in two or more subtasks, in a coordinated way. The distributive property is a good example of vertical complexity. Let's take the following example: $a \times (b + c) = (a \times b) + (a \times c)$. In order to correctly perform the task, one should multiply the element *a* by *b* and by *c*, separately, and then sum the results, or sum *b* with *c*, and then multiply by *a*. If someone change the order of execution of the actions, e.g. ($a \times b$) + c, the result won't be right. So, requires the two actions of addition and multiplication to be performed in a certain order, thus, coordinated.

Formally, one task is more hierarchically complex than another task if all of the following are true.

a) It is defined in terms of two or more lower-order task actions. In mathematical terms, this is the same as a set being formed out of elements. This creates the hierarchy.

   i. A = {a, b}, where *a* and *b* are "lower" than A and compose the set A;

   ii. A ≠ {A,...}, where the A set cannot contain itself. This means that higher order tasks cannot be reduced to lower order ones. For example, postformal task actions cannot be reduced to formal ones.

b) It organizes lower order task actions. In mathematics' simplest terms, this is a relation on actions. The relations are order relations:

   i. A = (a, b) = {a, {b}} an ordered pair

c) This organization is non-arbitrary. This means that there is a match between the model that designates orders and the real world orders. This can be written as: Not P(a,b), not all permutations are allowed (see Commons & Pekker, 2008).

Briefly summarizing, the MHC postulates that actions at a higher order of hierarchical complexity: 1) are defined in terms of two, or more, lower-order actions; 2) organize and transform those actions, not just combine them in a chain; and 3) produce organizations of lower-order actions that are new and not arbitrary. The first two are also Piagetian postulates, but the third is not. The order of hierarchical (or vertical) complexity refers to the number of recursions that the coordinating actions must perform on a set of primary elements (Commons, 2008).

Commons and Pekker (2008), after presenting the formal description of the theory and demonstrating its axioms, showed its four consequences:

1) Discreteness: The order of hierarchical complexity (*h*) of any action is a nonnegative integer, presenting gaps between orders.

2) Existence: If there exists an action of order *n* and an action of order *n+2*, then there necessarily exists an action of order *n+1*;

3) Comparison: The orders of hierarchical complexity of any two actions can be compared. For any two actions *A* and *B*: $h(A) > h(B)$ or $h(A) < h(B)$ or $h(A) = h(B)$.

4) Transitivity: For any three actions *A, B* and *C,* if $h(A) > h(B)$ and $h(B) > h(C)$, then $h(A) > h(C)$.

Because hierarchical complexity is a property of tasks, performance is separated from tasks. Stage is defined as the most hierarchically complex task solved. Each task that occurs in a separate domain is considered separately. There is no structure of the whole, so in the DST, *decalâge* is the normal modal state of affairs.

Since the MHC is related to the phenomenon of difficulty, it has a broad range of applicability. The mathematical foundation of the model makes it an excellent research tool to be used by anyone examining performance that is organized into stages. It is designed simply to assess development based on the order of complexity which the individual utilizes to organize information. The MHC offers a singular mathematical method of measuring stages in any domain because the tasks presented can contain any kind of information. The model thus allows for a standard quantitative analysis of developmental complexity in any cultural setting. Other advantages of this model include its avoidance of mentalistic or contextual explanations, as well as its use of purely quantitative principles which are universally applicable in any context. Cross-cultural developmentalists and animal developmentalists; evolutionary psychologists, organizational psychologists, and developmental political psychologists; learning theorists, perception researchers, and history of science historians; as well as educators, therapists, and anthropologists can use the MHC to quantitatively assess developmental stages.

The development of metrics in developmental psychology has been one of the challenges and needs of the area (Van Geert & Steenbeek, 2005; Fischer & Rose, 1999), and is considered crucial in guiding research and professional practice (Stein & Heikkinen, 2009). The Hierarchical Complexity Score System – HCSS (Commons, LoCicero, Ross & Miller, 2010); Dawson, Commons, Wilson, & Fischer, 2005) and the Lectical Assessment System – LAS (Dawson-Tunik, 2004) represent general, reliable, valid, domain-free scales or metrics (Dawson, 2004). These metrics were studied by Dawson (2000, 2001, 2002, 2003, 2004) who compared them with domain-specific scales, such as the *Good Life Scoring System* (Armon, 1984), the *Standard Issue Scoring System* (Colby & Kohlberg, 1987a,b) and the *Perry Scoring System*

(Perry, 1970). Dawson (2003) points out that, in spite of measuring the same latent variable, the domain-free scales present better internal consistency, allow meaningful comparisons across domains and contexts, and enable the examination of the relationship between developmental stages and conceptual content. Moreover, the HCSS and the LAS are considered two of few *calibrated developmental metrics* in use, being studied in terms of their construct and congruent validity, internal consistency and inter-rater reliability, providing evidence~~s~~ of fine grained interval scales (Stein & Heikkinen, 2009).

Despite the importance in guiding developmental and psycho-educational research and practice, the domain-specific scales demand various trained scoring analysts, with high agreement between them, require a considerable time for large scale evaluation and are vulnerable to subjective bias. So, the construction of objective large-scale tests can help the field to move beyond these challenges, bringing speed and lower cost-procedures for evaluating discontinuities.

As argued before, the MHC can be used not only to construct analytic scales, but also for the construction and design of tests, tasks and vignettes. Tasks have been created in a number of domains, based on the MHC or DST (as seen in Table 1).

Table 1.

Some Instruments Based on the Model of Hierarchical Complexity and/or Dynamic Skill Theory

PROBLEM-SOLVING
Algebra (Richardson & Commons, 2008)
Balance Beam (Dawson, Goodheart, Draney, Wilson, & Commons, 2010)
Infinity (Mathematics) (Richardson & Commons, 2008)
The Laundry Problems (Goodheart & Dawson, 1996; Goodheart, Dawson, Draney, & Commons, 1997)
The Combustion Problem (Bernholt, Parchmann, & Commons, 2008).
VIGNETTES
Social perspective-taking (Commons & Rodriguez, 1990; 1993)
Informed consent (Commons & Rodriguez, 1990, 1993; Commons, Rodriguez, Adams, Goodheart, Gutheil, & Cyr, 2006)
Attachment and Loss (Miller & Lee, 2000)
Workplace organization (Bowman, 1996a; 1996b)
Workplace culture (Commons, Krause, Fayer, & Meaney, 1993)
Political development (Sonnert & Commons, 1994)

Relationships (Armon, 1984a)
Views of the "good life" (Danaher, 1993; Dawson, 2000; Lam, 1994)
Epistemology (Kitchener & King, 1990; Kitchener & Fischer, 1990)
Moral Judgment (Armon & Dawson, 1997; Dawson, 2000)
The Helper-Person Problem, The Incest Dilemma Against, The Pro-Death Penalty Dilemma, The Anti-Death Penalty Dilemma, The Politician-Voter Problem, The Christ Stoning Case Without Sin (Miller, Bett, Ost, Commons, Day, Robinett, Ross, Marchand, & Lins, 2008)
OTHER
Four Story problem (Commons, Richards & Kuhn, 1982; Kallio & Helkama, 1991)
Counselor stages (Lovell, 2002)
Loevinger's Sentence Completion task (Cook-Greuter, 1990)
Report patient's prior crimes (Commons, Lee, Gutheil, Goldman, Rubin, Appelbaum, 1995)
Causing religious beliefs / Causing atheism (Miller, Harrigan, Commons, & Commons-Miller, 2008)
The Student-Bully Problem (Joaquim, 2011)

Constructing calibrated tests for developmental stage identification requires a specific design that is defined by Commons and colleagues (Commons & Pekker, 2008; Commons newest axiom paper – This issue). This design involves: 1) grouping items with same hierarchical complexity $[h(\mathrm{i}_1) = h(\mathrm{i}_2)= h(\mathrm{i}_3) = ... \ h(\mathrm{i}_n)]$ within stages; and 2) using items with increasing hierarchical complexity $[h(\text{Stage } 1) < h(\text{Stage } 2) < h(\text{Stage } 3) < ... \ h(\text{Stage } k)]$ between stages. The first deals with item or task equivalence, important in order to avoid the elaboration of an anomalous scale that confuses its analysis (Fischer & Rose, 1999). The second makes possible the identification of discontinuous, stage-like development, with gaps between different orders. There is an expected item structure of any instrument construct based on the MHC. That structure focuses on both strategies in order to identify developmental stages should be as close as possible to the diagram below (Fig. 1). Each blue box in the Figure 1 represents a cluster of items of the same unidimensional domain. Within a single box, the items have the same Order of Hierarchical Complexity (h) in that domain. The OHC of the items increases from stage 1 (φ_1) to stage k (φ_k), so that $h(\varphi_1) < h\ (\varphi_2) < \cdots < h(\varphi_k)$ (Consequences 2, 3 and 4 of the formal MHC). Furthermore, the figure shows the expected gaps between the clusters of adjacent OCH items (see Figure 1).

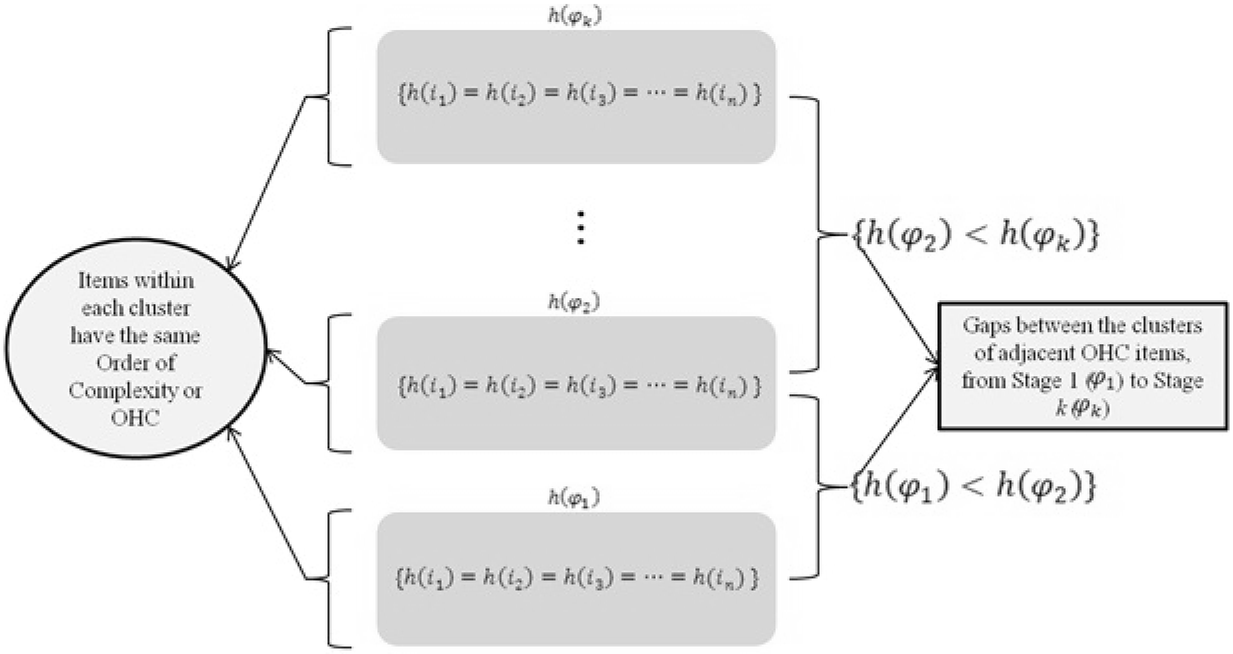

$h(\varphi_k)$
$\{h(i_1) = h(i_2) = h(i_3) = \cdots = h(i_n)\}$
$h(\varphi_2)$
$\{h(i_1) = h(i_2) = h(i_3) = \cdots = h(i_n)\}$
$h(\varphi_1)$
$\{h(i_1) = h(i_2) = h(i_3) = \cdots = h(i_n)\}$
Items within each cluster have the same Order of Complexity or OHC
$\{h(\varphi_2) < h(\varphi_k)\}$
$\{h(\varphi_1) < h(\varphi_2)\}$
Gaps between the clusters of adjacent OHC items, from Stage 1 (φ_1) to Stage k (φ_k)

Fig. 1 Expected Item Structure of instruments constructed focusing on the vertical complexity within a specific domain (unidimensional)

Beyond both strategies, a good measure or ruler needs to address a single trait or dimension, be constructed based upon an explicit theory or model of development (Stein, Dawson & Fischer, in press), be submitted to empirical investigation, aiming to test the expected equivalence and order of items, and determine other scale properties (Fischer & Dawson, 2002; Fischer & Rose, 1999). Commons and colleagues (Commons and Pekker, 2008; Commons newest axiom paper – get citation) evaluate the expected equivalence and order of items from the developmental test design through the Rasch family of models (Andrich, 1988; Rasch, 1960). The dichotomous Rasch Model (Rasch, 1960/1980), also called Simple Logistic Model (SLM) for dichotomous responses (Andrich, 1988), establishes that the right/wrong scored response X_{vi}, that emerges from the encounter between the person v and the item i, depending upon the performance β of that person and on the difficulty δ of the item. Its relation can be expressed as the following probabilistic function:

$$P\,\{X_{vi} = x\} = \frac{e^{x(\beta_v - \delta_i)}}{1+\,e^{(\beta_v - \delta_i)}} \qquad (1)$$

The Rasch model deals with the relationship between the person ability and item difficulty in a probabilistic way. Both parameters are allocated on a single abstract continuum that goes from "low" to "high" ("more" or "less", etc), concerning just one attribute of the object (or attitude, or behavior) measured, thus *unidimensional*. In the Classical Test Theory (CTT) the corresponding "parameter" for the Rasch's person performance (β_v) is the estimated *true score* ($\hat{T}_v$), or the score reported on test-score scale (normally distributed) (Hambleton & Jones, 1993). It can indicate the "position" of the person on the construct measured, but unlike the SLM, needs a representative sample for unbiased item estimates, a norm group for comparison between individuals, giving meaning to the scores, and a normally distributed score for achieving interval scales properties (Embreston & Reise, 2000).

Some authors argue that the dichotomous Rasch model is the simplest Item Response Theory model (one-paramenter model) (Bock & Jones, 1968; Hambleton, 2000). However, Andrich (2004) argues that differently from the traditional IRT paradigm, in which one chooses the model to be used (one, two or three parameters) according to which better accounts for the data, in the Rasch Paradigm "the SLM is used because it arises from a mathematical formalization of invariance which also turns out to be an operational criterion for fundamental measurement" (p.15). So, instead of data modeling, the Rasch's paradigm focuses on the verification of data fit to a fundamental measurement criterion, compatible with those found in the physical sciences (Andrich, 2004. p.15).

From among the benefits of using the Rasch family of models for measurement, some should be highlighted. In sum, it allows the construction of objective and additive scales, with equal-interval properties (Bond & Fox, 2001; Embreston & Reise, 2000), it produces linear measures, gives estimates of precision, allows the detection of lack of fit or misfit and enables the parameters' separation of the object being measured and of the measurement instrument (Panayides, Robinson & Tymms, 2010). It also makes possible the reduction of all of a test's items into a common developmental scale (Demetriou & Kyriakides, 2006), collapsing in the same latent dimension person's abilities and item's difficulty (Bond & Fox, 2001; Embreston & Reise, 2000; Glas, 2007), and enables the verification of hierarchical sequences of both item and person, being especially relevant to developmental stage identification (Dawson, Xie & Wilson, 2003).

Through the assumptions and procedures introduced by Commons and colleagues (Commons and Pekker, 2008; Commons newest axiom paper – get citation) it has become possible to design and construct valid and reliable developmental metrics, tests and tasks, bringing new empirical evidence that helps reveal stage-like discontinuous growth. Following this tradition, two exploratory studies about the construction, challenges and initial results from the construction of an objective, large-scale instrument, named the Inductive Reasoning Developmental Test (IRDT), developed by Gomes and Golino (2009). These studies will be presented in some detail with the aim of unpacking the challenges involved in the construction of a developmental test, and will present a methodology for developmental stage identification. This methodology is put forward as one of the moves that can help uncheck the idea of stages within the virtual game of "Developmental Chess", together with other moves published elsewhere (Demetriou & Kyriakides, 2006; Rijmen, De Boeck, & Van der Mass, 2005).

Study I: Uncovering Discontinuities, and Finding Alternative Sources of Difficulty Beyond Vertical Complexity

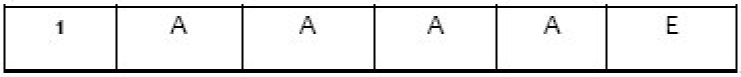
The purpose of Study 1 was to construct the initial version of the instrument, and in so doing, assess the scale structure of the items, verifying if they presented previously predicted orders and gaps, and to investigate the initial estimates of reliability and unidimensionality, among other scale properties, using Rasch analysis.

The Inductive Reasoning Developmental Test – IRDT (Gomes & Golino, 2009) is a pencil-and-paper instrument design to assess developmentally sequenced and hierarchically organized inductive reasoning. It is an extension, in terms of complexity, from the *Indução* test, which compose the fluid intelligence test kit (Gomes & Borges, 2009) of the Higher-Order Cognitive Factors Kit (Gomes, 2010). The domain of inductive reasoning was used because it is one of the best indicators of fluid intelligence (Carroll, 1993). The construction of the IRDT, from the original *Indução* items, is due to a larger challenge that concerns the construction of an intelligence battery to identify developmental stages.

The sequence of IRDT was constructed based on the MHC and on Fischer's Dynamic Skill Theory. It was designed to identify six developmental stages (or levels), that will be named based in both theories, respectively: Pre-operational or Single Representations (Pre-op/SR);

Primary or Representational Mappings (Prim/RM); Concrete or Representational Systems (Conc/RS); Abstract or Single Abstractions (Abst/SA); Formal or Abstract Mappings (Form/AM); and Systematic or Abstract Systems (Syst/AS). Each stage is composed of eight items with the same order of hierarchical complexity (OHC), for a total of forty-eight items. Each item is composed of four letters, or sequence of letters, with a specific rule (correct items), plus one letter or sequence with a different rule (exception). The task is to discover which letter or sequence is the exception. From stage to stage, there is a difference of +1 in the Order of Hierarchical Complexity (OHC). The instructions for performing the test is as follow: “You’ll be presented several reasoning tasks (items). In each task (item) you have five letters or sequence of letters. Among the five letters or sequence of letters, four of them have a specific rule, and one has a rule that is different from the others. Your challenge is to identify (marking with an X) the letter or the sequence of letters that has a different rule, compared to the other four. Each task (item) is displayed in a specific row, beginning with a number, from 1 to 48. You have no time limit. Solve as many tasks (items) as you can.”

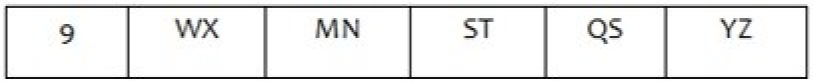
Pre-operational or Single Representations (Pre-op/SR): Each item is composed of specific letters. The rule is “equal letter”, and the exception is a different one. (see Figure 2)



1	A	A	A	A	E

Fig. 2 Example: Item 1, Stage Pre-op.

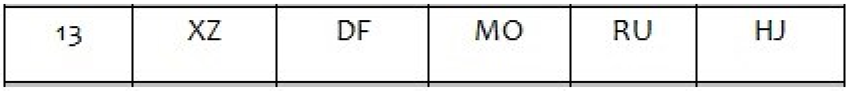
Primary or Representational Mappings (Prim/RM): Eight items were created for this stage. Four of them have a specific rule: there is no jump in the letters’ sequence. In the example below, the first option is composed of WX. There is no other letter between them, so they form a non-jump sequence (Rule 1). The exception, however, is a conjoint of two letters that jumps one letter of the alphabetic sequence (e.g. QS). (see Figure 3)



9	WX	MN	ST	QS	YZ

Fig. 3 Example: Item Prim/MR1 – Rule 1

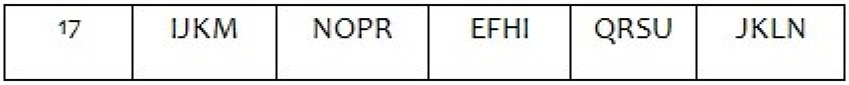
The other four items of the Primary Stage follows the same structure, but have different rules. The majority of the options jump one letter of the alphabetic sequence (Rule 2). So, in the example below, the option *DF* jumps the letter *E*. The exception is a conjoint of two letters that jumps two letters of the alphabetic sequence (e.g. RU) (see Figure 4).



13	XZ	DF	MO	RU	HJ

Fig. 4 Example: Item 13, Primary/MR – Rule 2

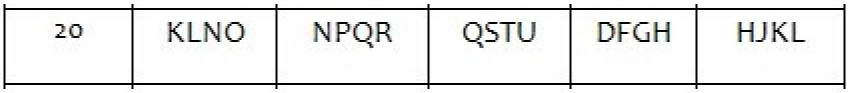
Concrete or Representational Systems (Conc/RS): All items are composed of four sets of four letters with one of the three following rules. In Rule 3 there is a jump of one letter only between the last two letters. For one example, see the item below. Between I and J, and between J and K, there is no other letter. However, there's a jump between K and M. The exception, in this item (17), is represented by the sequence EFHI, where the jump is located between the two letters in the middle (FH) (see Figure 5).



17	IJKM	NOPR	EFHI	QRSU	JKLN

Fig. 5 Example: Item 17, Concrete/RS – Rule 3

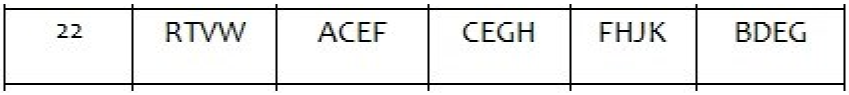
In Rule 4, the jump occurs between the first pair of letters, and the exception is the option where the jump occurs between the two middle letters. The example below shows item 20. Note that the option NPQR presents a jump between N and P, like three other options. However, the first option (KLNO) presents a jump between the two middle letters, i.e. L and N (see Figure 6).



20	KLNO	NPQR	QSTU	DFGH	HJKL

Fig. 6 Example: Item 20, Concrete/RS – Rule 4

Finally, in rule 5 the jump occurs twice, between the two first pairs of letters. In the exception, the jumps occur between the first pair and between the last pair of letters. See the example below. In item 22, in the first option (RTVW) there is a jump between R and T, and between T and V, as in three other options. However, in the option BDEG, the jumps occur between B and D, and E and G (see Figure 7).

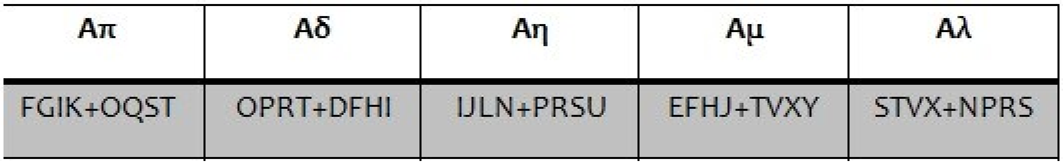


22	RTVW	ACEF	CEGH	FHJK	BDEG

Fig. 7 Example: Item 22, Concrete/RS – Rule 5

So, the first two items (Prim/RS1 and Prim/RS2) use rule 3, the items Prim/RS3 and Prim/RS4 use rule 4, and the other four items use rule 5.

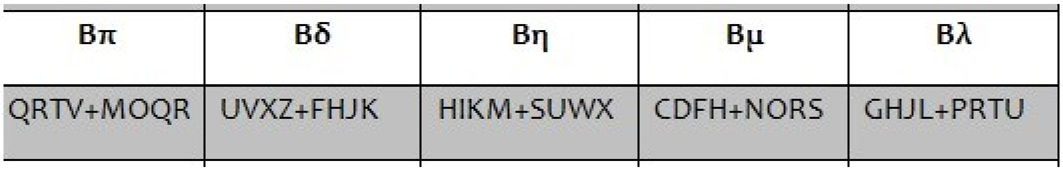
Abstract or Single Abstractions (SA): Different from all other stages, here a table is introduced with codes referring to a coordination of two sets of four letters, in which the rules and exceptions presented at the Concrete/SR's items are also coordinated, forming new rules and exceptions. This coordination is shown by the *plus* sign between the letter sequences (see Figure 8).



Aπ	Aδ	Aη	Aμ	Aλ
FGIK+OQST	OPRT+DFHI	IJLN+PRSU	EFHJ+TVXY	STVX+NPRS

Fig. 8 Example: Table Row 1, Abstract/SA

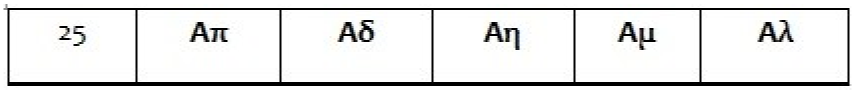
The table has eight code rows, each beginning with an alphabetic letter followed by a Greek letter. So, the first code row has letter A followed by different Greek letters, while the second code row has letter B followed by the same Greek letters, and so on (see Figure 9).



Bπ	Bδ	Bη	Bμ	Bλ
QRTV+MOQR	UVXZ+FHJK	HIKM+SUWX	CDFH+NORS	GHJL+PRTU

Fig. 9 Example: Table Row 2, Abstract/SA

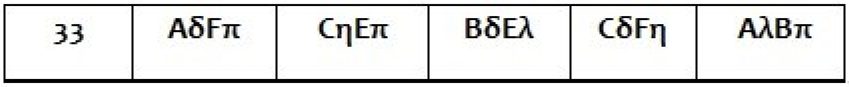
The item to be answered is composed only by the table codes, in sequence. For example see Figure 10.



25	Aπ	Aδ	Aη	Aμ	Aλ

Fig. 10 Example: Item 25, Abstract/SA

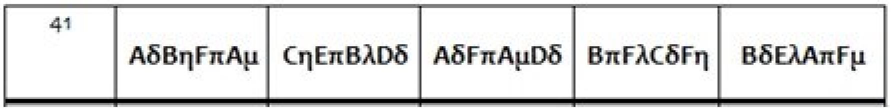
Formal or Abstract Mappings (Form/AM): All items are composed of a coordination of two codes, based on those presented at the Abstract Stage's table (see Figure 11).



33	AδFπ	CηEπ	BδEλ	CδFη	AλBπ

Fig. 11 Example: Item 33, Formal/AM

Systematic or Abstract Systems (AS): All items are composed by a set of four codes, based on the previous presented at Abstract Stage's table (see Figure 12).



41	AδBηFπAμ	CηEπBλDδ	AδFπAμDδ	BπFλCδFη	BδEλAπFμ

Fig. 12 Example: Item 41 ,Systematic/AS

All items of the same stage were presented together at a specific page, so different stages were in different pages. The alphabetic sequence (all letters from A to Z) were printed above the items in each page, for consultancy. The order of hierarchical complexity is represented in the figure 13 below. The Systematic items (OHC 11) coordinate two formal (OHC 10) components. By its turn, the formal items coordinate two abstract (OHC 9) components. The abstract items coordinate two concrete (OHC 8) components. The concrete items coordinate two primary (OHC 7) components. Finally, the primary items coordinate two pre-operational (OHC 6) components (see Figure 13).

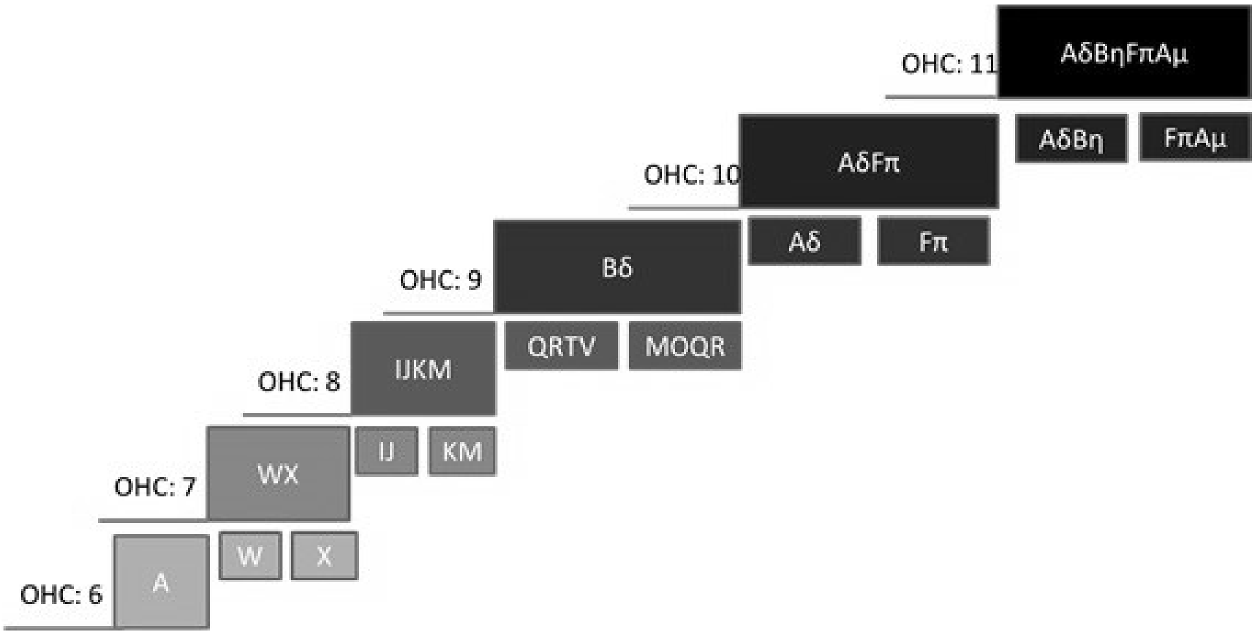


Fig. 13 Hierarchy of items

Method

Participants

In Study 1, the IRDT was administered to a convenience sample composed by 167 Brazilian people (50.3% men, 49.7% women) aged between 6 to 58 years (*M* = 18.90, *SD* = 9.70). The sample was intentionally broad, and had a distribution of 15.6% from 6 to 12 years, 27.5% from 13 to 15 years, 35.9% from 16 to 20 years, and 21% beyond 20 years. All the participants were from the city of Belo Horizonte, state of Minas Gerais, Brazil.

Procedure

The data were collect by the first author and by thirty Psychology undergraduate students, enrolled in a first semester Cognitive Development class, the latter of whom were trained in how to administer the instrument properly. The author first administered the instrument to the undergraduate students (whose data are being used in this analysis), and to 47 first year high school students from a public school. Each undergraduate student was assigned ~~t~~o administer the IRDT to three different people from 6 to 60 years of age. Participation was voluntary, with participants agreeing to participate after the purpose of the study was explained. They were informed that their answers would be kept confidential, and that all procedures guaranteeing the privacy of their results would be adopted. They then signed an inform consent form, as required by the guidelines of the Ethical Committee of the Universidade Federal de Minas Gerais, Brazil.

Data Analysis

In the first part of the data analysis the dichotomous Rasch Model is used, making it possible to reduce the items from the IRDT into a developmental scale (Demetriou & Kyriakides, 2006), collapsing at the same level person's abilities and item's difficulty (Bond & Fox, 2001; Embreston & Reise, 2000; Glas, 2007). It also enables the verification of hierarchical sequences of both item and person, being especially relevant to developmental stage identification (Dawson, Xie & Wilson, 2003).

To verify the adjustment of the data to the model, the Infit (information-weighted fit) mean-square statistic is used. It represents "the amount of distortion of the measurement system" (Linacre, 2002. p.1). Values between 0.5 and 1.5 logits are considered productive for measurement, and <0.5 and between 1.5 and 2.0 are not productive for measurement, but do not degrade it (Wright & Linacre, 1994). The unidimensionality of the instrument can be checked by a number of procedures, each one complementing the other (see Tennant & Pallant, 2006). Here, unidimensionality will be addressed using only the model fit statistics – i.e. if the data fit the model, one of the consequences is the linearity of the measure, its unidimensionality, and so on – and the principal contrast, which can be verified through the percentage of variance explained by measures, and by the percentage of unexplained variance in the first contrast. The former should be closer to or greater than 60% (Peeters & Stone, 2009), while the latter should be closer to or less than 10%.

In the second part of the analysis, the spacing of Rasch scores between items of adjacent orders of hierarchical complexity is described. The Rasch scores represent the difficulty of an item (δ), which is its location at the latent variable continuum. It would have been good to compare the Rasch Scores for every item from adjacent orders of hierarchical complexity, but because there were so many items, this would have produced too many comparisons. To reduce the number of comparison pairs, each item's Rasch score was subtracted from the mean Rasch score of the items from the next higher order of complexity. This calculation is represented by the Formula 2:

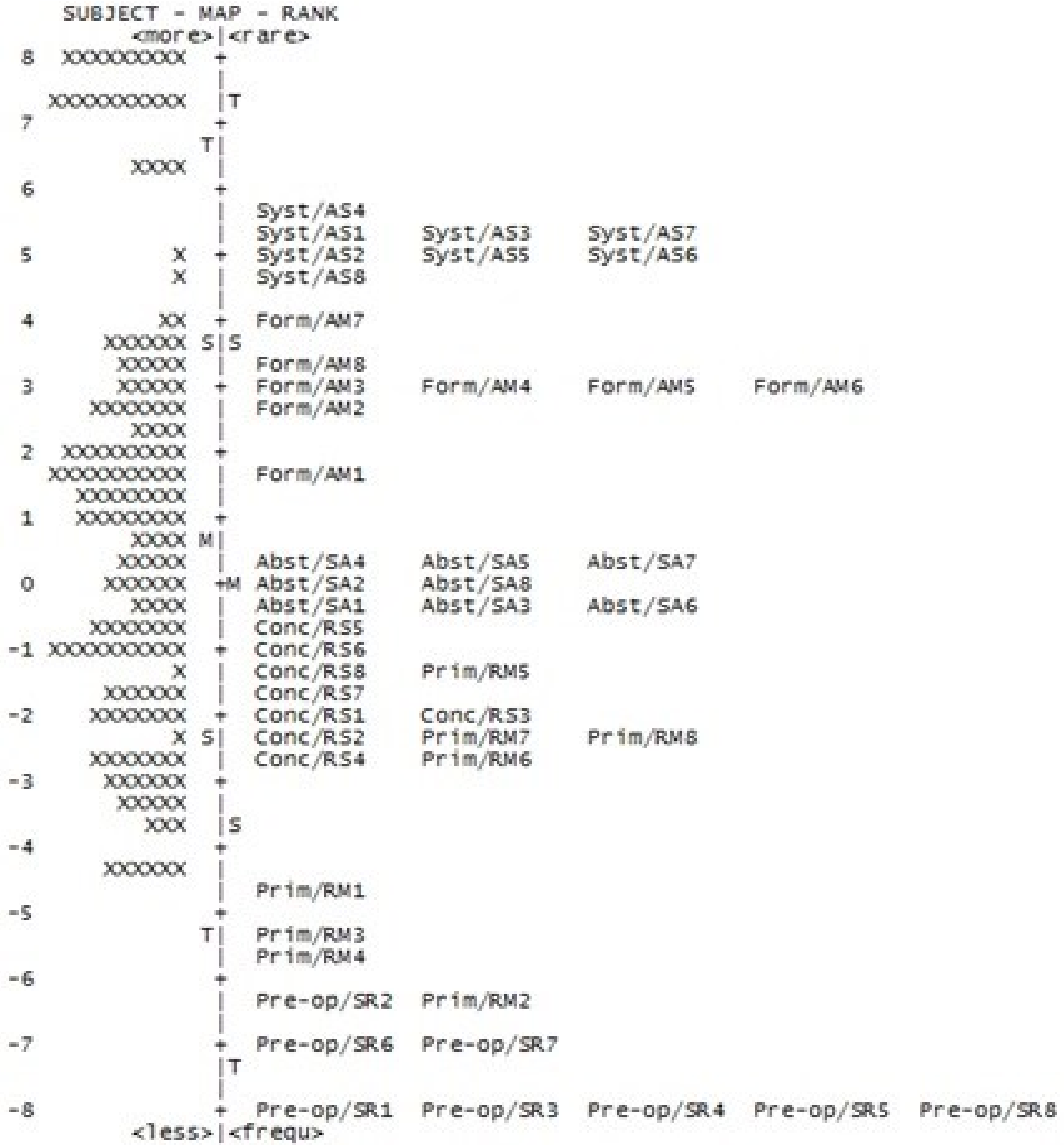
$$\overline{\mathrm{X}}_{k+1} - \delta_{i_k} = Adj\delta_{i_k} \qquad (2)$$

where $\overline{\mathrm{X}}_{k+1}$ is the mean of the next higher order of complexity (or Stage k+1), and δ_{i_h} is the difficulty of item *i* from order *k* (or Stage k) , producing the adjusted difficulty of item *i.* To verify if the differences between difficulties of items from order *k* and the mean difficulty of the order k+1 are statistically significant, the One-Sample t-test is used, with a 95% confidence interval. The effect size is calculated using the Cohen's *d*.

Results

The Rasch dichotomous model (Andrich, 1988; Rasch, 1960) was calculated using the software Winsteps (Linacre, 1999, 2011). Out of the 48 items, 5 were responded to correctly by all participants (Pre-op/SR1, Pre-op/SR3, Pre-op/SR4, Pre-op/SR5 and Pre-op/SR8). The reliability for the forty-three non-extreme items was .99, and for the full scale (48 items) the reliability was .97. The Infit mean was .87 (*SD* = .28; *Max* = 1.69; *Min* = .39), falling within the acceptable fit range. The person reliability was .95, which is estimated to indicate the degree to which a person's response pattern conforms to the difficulty structure of the measure (Hibbard, Collins, Mahoney & Baker, 2009). The principal contrast showed that the raw variance explained by measures (modeled) is 70.6%, and that the unexplained variance in the first contrast (modeled) is 10.4%, suggesting that the instrument can be thought of as unidimensional.

The variable map (Figure 2) illustrates the scale for the 48 items of the IRDT with item difficulties (on the right) and person measures (on the left) calibrated on the same scale. It is visually possible to identify clear item clusters in the Systematic/Abstract Systems' stage (Syst/AS1, Syst/AS2, Syst/AS3, …, Syst/AS8) and in the Formal/Abstract Mappings's stage (Form/AM1, Form/AM2, Form/AM3, …, Form/AM8), with a gap between them. The Abstract/ Single Abstraction's items presented a cluster (they are all together without any other stage's items), but did not present a gap in relation to the Concrete/Representational System's items. Some Primary/Representational Mapping's items (Prim/RM5, Prim/RM6, Prim/RM7, Prim/RM8), had difficulties very close to the Concrete/RS's items, making one big item set. The other Primary/RM's items (i.e. Prim/RM1, Prim/RM2, Prim/RM3 and Prim/RM4) were less difficult than other items of the same stage. Moreover, they presented a gap in relation to the item's set composed by the other Primary items and by the Concrete ones. Finally, the relative position of person (left) and item (right), shows the IRDT as an easy test for 23 participants (*Mean ability* = 7.66, *SD* = 0.81). The whole sample mean ability was 1.15 with standard deviation of 3.40 logits (see Figure 14).



```
    SUBJECT - MAP - RANK
         <more>|<rare>
 8  XXXXXXXXX  +
               |
   XXXXXXXXXX  |T
 7             +
              T|
         XXXX  |
 6             +
               |  Syst/AS4
               |  Syst/AS1    Syst/AS3    Syst/AS7
 5          X  +  Syst/AS2    Syst/AS5    Syst/AS6
            X  |  Syst/AS8
               |
 4         XX  +  Form/AM7
       XXXXXX S|S
        XXXXX  |  Form/AM8
 3      XXXXX  +  Form/AM3    Form/AM4    Form/AM5    Form/AM6
      XXXXXXX  |  Form/AM2
         XXXX  |
 2  XXXXXXXXX  +
   XXXXXXXXXX  |  Form/AM1
     XXXXXXXX  |
 1   XXXXXXXX  +
         XXXX M|
        XXXXX  |  Abst/SA4    Abst/SA5    Abst/SA7
 0     XXXXXX  +M Abst/SA2    Abst/SA8
         XXXX  |  Abst/SA1    Abst/SA3    Abst/SA6
      XXXXXXX  |  Conc/RS5
-1 XXXXXXXXXX  +  Conc/RS6
            X  |  Conc/RS8    Prim/RM5
       XXXXXX  |  Conc/RS7
-2    XXXXXXX  +  Conc/RS1    Conc/RS3
            X S|  Conc/RS2    Prim/RM7    Prim/RM8
      XXXXXXX  |  Conc/RS4    Prim/RM6
-3     XXXXXX  +
        XXXXX  |
          XXX  |S
-4             +
       XXXXXX  |
               |  Prim/RM1
-5             +
              T|  Prim/RM3
               |  Prim/RM4
-6             +
               |  Pre-op/SR2  Prim/RM2
               |
-7             +  Pre-op/SR6  Pre-op/SR7
               |T
               |
-8             +  Pre-op/SR1  Pre-op/SR3  Pre-op/SR4  Pre-op/SR5  Pre-op/SR8
         <less>|<frequ>
```

Fig. 14 Variable Map showing the IRDT's items

The One-Sample t-test, with 95% confidence interval, shows that the comparisons of difficulty between Pre-operational and Primary, Primary and Concrete, Concrete and Abstract, Abstract and Formal, and between Formal and Systematic were significant. Moreover, the effect sizes (d') were large (see Table 2).

Table 2

One-sample *t*-tests of Mean Item Difficulties for Different OHC's

Stages	Test Value = 0							
						95% Confidence Interval of the Difference		
	t	DF	Sig. (2-tailed)	Mean Difference	Std. Deviation	Lower	Upper	Effect Size (d')
Pre-op/SR and Primary/RM	13,58	7	0,00	3,82	0,80	3,15	4,48	4,80
Primary/RM and Concrete/RS	3,29	7	0,01	2,18	1,87	0,61	3,74	1,16
Concrete/RS and Abstract/SA	7,99	7	0,00	1,69	0,60	1,19	2,18	2,82
Abstract/AS and Formal/AM	36,01	7	0,00	2,89	0,23	2,70	3,08	12,73
Formal/AM and Systematic/AS.	9,49	7	0,00	2,28	0,68	1,71	2,85	3,35

Discussion

The current study aimed to assess the scale structure of the items, verifying whether they represented previously predicted orders and gaps (see Fig.1), and to investigate the initial estimates of reliability and unidimensionality, among other scales properties, using Rasch analysis. The result suggests the unidimensionality of the items, to some extent, since the percentage of raw variance explained by the measures (modeled) is moderately high (70.6%), and the principal components analysis of the residuals gave an unexplained variance of 10.4% for the first contrast. The items' adjustment to the model was verified through the Infit index, which was found to have a mean of .87 and a standard deviation of .28. The minimal Infit value was .39 (Item System/AS4) and the maximum was 1.69 (Item Primary/MR5), and all other non-extreme items had Infits smaller than 1.32. This is considered to reflect a good fit to the model. The person and item reliabilities were good (.97 and .95, respectively). After assessing some of the psychometric properties of the measures, it was necessary to look more closely at the variable map (Fig.14).

The Pre-operational/Single Representation stage presented two sets of item difficulties, i.e. items Pre-op/SR1, Pre-op/SR3, Pre-op/SR4, Pre-op/SR5 and Pre-op/SR8 were shown to be less difficult than items Pre-op/SR2, Pre-op/SR6 and Pre-op/SR7. This gap between items with the same predicted OHC suggests that there was a problem in designing these items. One hypothesis to explain this effect could be that they are more horizontally complex. The Preo-operational items are composed of four equal letters plus a different letter, requiring the participant only to discriminate a set of five simple stimuli, choosing the dissimilar one. The items Pre-op/SR2, Pre-op/SR6 and Pre-op/SR7 may have been more difficult because the letters provided as options, in each item, were closer in graphical terms. The item Pre-op/SR2, for example, was composed by four "O" and one "Q". The visual stimuli of both letters are graphically closer, differing by the little "dash" on the bottom of Q. Previous research has shown that the structure of cognitive processing is composed of cascade-like relations (Demetriou, Christou, Spanoudis, & Platsidou, 2002; Demetriou, Mouyi, & Spanoudis, 2008) between processes with increasing complexity, beginning with speed processing (the most basic component of the cognitive architecture), followed by perceptual discrimination, perceptual control, conceptual control, short-term memory, working memory and, finally, reasoning processes. According to Demetriou, Mouyi and Spanoudis (2008), perceptual discrimination "reflects sheer speed of processing together with the processes required to discriminate between two simple stimuli and identify the target one" (p. 439). So, when comparing different stimuli, those whose difference are based on small tiny cues (e.g. the little dash of letter Q), demand a higher perceptual discrimination than those having more cues (e.g. comparing "A" with "E"). Thus, Pre-op/SR2, Pre-op/SR6 and Pre-op/SR7 are more *horizontally* complex than the other four Pre-operational items, because they demand a slight higher level of perceptual discrimination. In sum, it seems that in items from the Pre-operational order it is important to control as much as possible the perceptual discrimination required for the item or task, in order to avoid interference from the standpoint of horizontal complexity.

The next order's items also present two set of difficulties. The items Prim/RM1, Prim/RM2, Prim/RM3 and Prim/RM4 were the easiest items of the Primary stage, probably because they were constructed according to the Rule 1, i.e. four options with no jump between the pair of letters, and one option jumping one letter. The other four Primary items where constructed according to the Rule 2, which states a jump of one letter between each pair of letters (4 options), and one option jumping two letters. Our hypothesis is that when dealing with items constructed according to Rule 2, the participants needed to store and deal with more information in Working Memory (Demetriou et al., 2002, 2008; Pascual-Leone, 1984), which could horizontally increase the complexity of the task. A similar effect also seems to occur with the next order's items. Note the items Conc/RS5, Conc/RS6, Conc/RS7 and Conc/RS8, which are the most difficult concrete items, have a mean difference of .92 logits from the Conc/RS1, Conc/RS2, Conc/RS3 and Conc/RS4. This might be because the most difficult items have a rule which involves one more bit of information, being more horizontally complex than the items Conc/RS1, Conc/RS2, Conc/RS3 and Conc/RS4. Originally, we varied some of the rules somewhat in order to make the task less boring, and to avoid possible fatigue from the repetition of procedures employed to answer an item or task. However, our result suggests that changing some items' rules within a certain OHC can compromise the quality of the stage identification. It seems that a good strategy for developmental test construction is trying always to elaborate items with the same rule within a single OHC.

The items from the Abstract, Formal and Systematic orders, on the other hand, are forming groups, or clusters, reflecting the fact that items within each are of the same hierarchical complexity (and are therefore grouped together), and items across each order are appropriately separated. The Abstract items, however, are not well separated from the Concretes items. It can be speculated that the way the tables of the Abstract order were constructed, having eight code rows, each beginning with an alphabetic letter followed by a Greek letter, decreases the difficulty of the items. The options of the items are all organized and well structured, and this organization seems to work as a support for the respondents.

In spite of providing good indicators of the items' structure, and enabling the verification of visual clusters of items, the Rasch analysis did not provide information regarding the size of the gaps between adjacent OHC. The one-sample t-tests, calculated for this purpose, showed that the differences between adjusted difficulties of items from adjacent orders are statistically significant, with large effect sizes. This provides some additional evidence that helps support the existence of developmental stages of inductive reasoning. However, this result should be carefully interpreted, and future studies should employ a more balanced sample, from childhood to adulthood.

Study II: Refining the IRDT and investigating its Construct/congruent Validity.

Study 2 aims to modify some items of the IRDT, based on the results from the first study, and, using Rasch analysis, assess its new scale structure, verifying whether the previously predicted orders and gaps, as well as the scale's reliability and unidimensionality.

Part I: Instrument improvement

From the results of Study I, we've modified some items of the IRDT. Basically, the modifications can be synthesized as follows. From the original eight Pre-operational items, those demanding high perceptual discrimination were excluded, due to close similarities and low graphical clues (such as Q and O, etc), except one. We left one item to verify whether it still has more difficulties than the other Pre-operational items. The others were all modified in order to obtain items with easily discriminative options, such as "R F F F F" (Item Pre-op/SR3) and "H H L H H" (Item Pre-op/SR8). At the Primary order we removed those items constructed based on Rule 2, in which the pair of letters jumps one letter of the alphabetic sequence, and replaced them with items constructed based on Rule 1, i.e. with no jump in the letters' sequence, except for the option that is the exception and therefore is correctly supposed to be chosen by the participants because it does not follow the rule. Finally, the last change in the instrument occurred with the Abstract items, more precisely in the tables where the coordination of Concrete sequences are displayed. Instead of having a specific alphabetic letter in each row, and a specific Greek letter in each column, forming a code composed by two symbols for each cell that contains a coordination of two Concrete sequences, the table was modified to contain only one symbol (Greek letter) per cell. Moreover, the Abstract items are now formed by options that are spread throughout the table, so the participant needs to locate each one, and try to figure out which has a coordination rule that differs from the other 4 options. In the first version of the IRDT, the Abstract items' options were organized in each row. Also, the "plus" (+) symbol that mediated the coordination of the two Concrete sequences was taken out. The other two orders' items remained the same, since they demand the coordination of actions from the previous adjacent OHC. In sum, we’ve remodeled the items within each order, focusing on its vertical complexity. Our hypothesis is that this “*verticalization*” provides a better stage identification, with visual clusters of items and gaps between adjacent OHC more clearly defined.

Method

Participants

In Study 2, the revised IRDT were administered to a convenience sample composed of 188 Brazilian people (42.3% men, 57.7% women) aged between 6 to 65 years (*M* = 21.45, *SD* = 14.31). The sample, again, was intentionally broad and had a distribution of 34.4% from 6 to 12 years, 13.4% from 13 to 15 years, 7.5% from 16 to 21 years, and 44.6% older than 21 years. All the participants were from the city of Belo Horizonte, state of Minas Gerais.

Procedure

The data were collect by the first author and by twenty five Psychology undergraduate students, enrolled in a second semester Cognitive Development class, who were trained to administer the instrument properly. The author first administered the instrument to the undergraduate students (and those which data are actually being used in this analysis). Each undergraduate student had to administer the IRDT to different people from 6 to 65 years old. Participation was voluntary. The potential participants had the purpose of the study explained to them. They were informed that their answers would be kept confidential, and that all procedures guaranteeing the privacy of their results would be adopted. They signed a inform consent, according to the guidelines of the Ethical Committee of the Universidade Federal de Minas Gerais, Brazil.

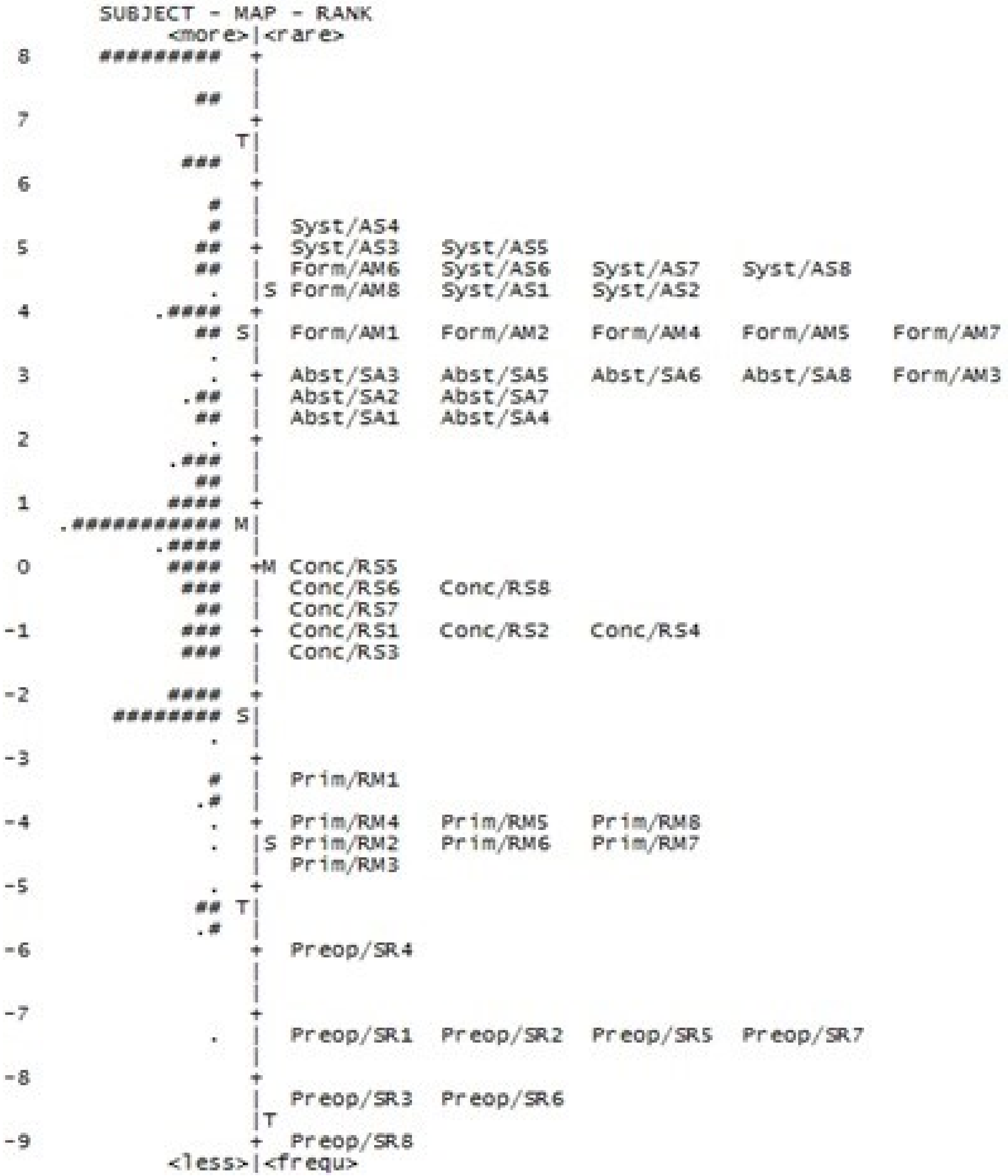
Data Analysis

The same data analytic process presented in Study 1 was adopted here. To assess the new scale structure of the IRDT, verifying if it presents the predicted orders and gaps, as well as its reliability and unidimensionality, we’ve employed the dichotomous Rasch model. To verify if the differences between the mean difficulty of items from order *k* and the mean difficulty of items from order *k+1* are statistically significant, the one-sample t-test is used, with 95% confidence interval. The effect size is calculated using Cohen’s d.

Results

The Rasch dichotomous model (Andrich, 1988; Rasch, 1960) was calculated using the software Winsteps (Linacre, 1999, 2011). From 48 items, only one was correctly responded to by all participants (Pre-op/SR8). The reliability for the full scale was .99, and its Infit mean was .94 (*SD* = .22; *Max* = 1.46; *Min* = .56). The person reliability was .95, which is estimated to indicate the degree to which a person's response pattern conforms to the difficulty structure of the measure (Hibbard, Collins, Mahoney & Baker, 2009). The principal contrast showed that the raw variance explained by measures (modeled) was 74.8%, and that the unexplained variance in the first contrast (modeled) was 12.9%, suggesting that the instrument can be thought of as unidimensional, even though the variance explained by the first contrast is higher than 10%. We argue that the variance explained by measures (modeled) is high enough to sustain its unidimensionality.

The variable map (Figure 15) illustrates the scale for the 48 items of the IRDT with item difficulties (on the right) and person (student) measures (on the left) calibrated on the same scale. It's visually possible to identify clear item clusters for almost all the orders, with a gap between them. However, two formal items, Form/AM6 and Form/AM8 had their scaled difficulties closer to the Systematic items, and one additional formal item, Form/AM3, had its scaled difficulty closer to the Abstract items. The only other difficulties were with the Pre-operational items, which were very spread out, but were nevertheless separated from the Primary items. Regarding the relative position of person (left) and item (right), the variable map shows the IRDT was an easy test for 28 participants (*Mean ability* = 7.86, *SD* = 0.87). The whole-sample mean ability was 1.15 with standard deviation of 3.40 logits (see Figure 15).



```
           SUBJECT - MAP - RANK
                 <more>|<rare>
  8      #########  +
                    |
                ##  |
  7                 +
                   T|
               ###  |
  6                 +
                 #  |
                 #  |   Syst/AS4
  5             ##  +   Syst/AS3   Syst/AS5
                ##  |   Form/AM6   Syst/AS6   Syst/AS7   Syst/AS8
                 .  |S  Form/AM8   Syst/AS1   Syst/AS2
  4          .####  +
                ## S|   Form/AM1   Form/AM2   Form/AM4   Form/AM5   Form/AM7
                 .  |
  3              .  +   Abst/SA3   Abst/SA5   Abst/SA6   Abst/SA8   Form/AM3
               .##  |   Abst/SA2   Abst/SA7
                ##  |   Abst/SA1   Abst/SA4
  2              .  +
              .###  |
                ##  |
  1           ####  +
      .########### M|
             .####  |
  0           ####  +M  Conc/RS5
               ###  |   Conc/RS6   Conc/RS8
                ##  |   Conc/RS7
 -1            ###  +   Conc/RS1   Conc/RS2   Conc/RS4
               ###  |   Conc/RS3
                    |
 -2           ####  +
          ######## S|
                 .  |
 -3                 +
                 #  |   Prim/RM1
                .#  |
 -4              .  +   Prim/RM4   Prim/RM5   Prim/RM8
                 .  |S  Prim/RM2   Prim/RM6   Prim/RM7
                    |   Prim/RM3
 -5              .  +
                ## T|
                .#  |
 -6                 +   Preop/SR4
                    |
                    |
 -7                 +
                 .  |   Preop/SR1  Preop/SR2  Preop/SR5  Preop/SR7
                    |
 -8                 +
                    |   Preop/SR3  Preop/SR6
                    |T
 -9                 +   Preop/SR8
              <less>|<frequ>
```

Fig. 15 Variable Map showing the IRDT 2nd version's items

The one-sample t-test, with 95% confidence interval, shows that the comparisons between Pre-operational and Primary, Primary and Concrete, Concrete and Abstract, Abstract and Formal, and between Formal and Systematic were significant. Moreover, the effect sizes (d’) were large (see Table 3).

Table I

One-Sample T Test

Stages	Test Value = 0							
						95% Confidence Interval of the Difference		
	t	DF	Sig. (2-tailed)	Mean Difference	Std. Deviation	Lower	Upper	Effect Size (d’)
Pre-op/SR and Primary/RM	10,36	7,00	,00	3,61	,99	2,79	4,43	3,66
Primary/RM and Concrete/RS	22,94	7,00	,00	3,42	,42	3,06	3,77	8,11
Concrete/RS and Abstract/SA	23,03	7,00	,00	3,33	,41	2,99	3,67	8,14
Abstract/AS and Formal/AM	10,96	7,00	,00	1,14	,29	,89	1,38	3,87
Formal/AM and Systematic/AS.	4,78	7,00	,00	,88	,52	,44	1,31	1,69

Discussion

The evidence shows that modifying the IRDT, in order to eliminate some sources of horizontal complexity, produced an item structure closer to what was expected when constructing an instrument according to the MHC and using the strategies presented in the introduction (see Figure 1). In each OHC, the items are grouped forming a visual cluster, and presenting a gap in relation to the adjacent orders. Two Formal items had difficulties higher than expected (Form/AM6 and Form/AM8) and one was less difficult than predicted. However, this small deviation does not interfere with the spacing of its Rasch scores in relation to the adjacent orders of hierarchical complexity. The Pre-operational items have its scaled difficulties somewhat scattered through the less difficult end of the scale, an unexpected result to some extent, since the items were modified to contain stimuli that were expected to be easily discriminated (having many graphical clues). However, it can be speculated that the differences in difficulty of these items are due to factors other than the nature of each stimulus’ contribution to the increase in its horizontal complexity. In any case, the item Pre-op/SR4 presents a difficulty at least 1.26 logits higher than the other Pre-operational items. This result was expected, since the Pre-op/SR4 ("U U V U U") is the same in both versions of the IRDT, and presents options graphically close to each other, demanding a higher amount of perceptual discrimination.

Regarding the data's fit to the model, the modified version of the IRDT produced a better Infit mean of the items (.94), representing an increase of .06 over the items' Infit of the first version (.88). The percentage of variance explained by the measures also increased from 70.6 with the previous version to 74.8 with the new one. It can be speculated that when we eliminated part of the horizontal complexity of the items, the amount of variance explained by the unidimensional measure increased. So, the "verticalization" process seems to contribute to the measure, not only in terms of the theory behind the items, i.e. the Model of Hierarchical Complexity, and by consequence the expected item structure, but also in terms of the adjustment of the items to the model and to the amount of variance explained.

Now that the item structure is closer to the expected (Figure 1), and the items' fits are more adequate, it seems to be relevant to coordinate the Rasch metrics and the Orders of Hierarchical Complexity in a mathematical fashion, to obtain a score representing stage of performance. There is no direct way to obtain a person score that represents stage of performance from the estimates obtained through the Rasch Dichotomous model. This seems to be a dilemma, mainly because there is a difference in formal measurement theory terms between the OHC and the Rasch scores. The former is an analytic measure represented in an ordinal scale, while the latter are an empirical conjoint-interval measure. But, there's a way to calculate stage of performance from the Rasch estimates. It can be calculated only because the items have the properties previously expected, i.e. they form clusters or groups within each OHC, present significant gaps with higher effect size between adjacent orders, and have adequate fit to the Rasch model. So, meeting these conditions, one can apply the below formula:

$$\varphi_j = \frac{\beta_j - \overline{X}_k}{\overline{X}_{k+1} - \overline{X}_k} + OHC_k \qquad (3)$$

where φ_j is the stage of performance of person j, β is the Rasch score of that person, $\overline{X}_k$ is the mean difficulty of items on order k, $\overline{X}_{k+1}$ is the mean difficulty of items on the next adjacent order, and OHC_k is the number that represents the order of hierarchical complexity k. For computing the stage scores of people whose ability lies on the highest order measured, one needs to leave the denominator as $\overline{X}_k$. After computing the stage of performance for each person, it is possible to verify how well the stage scores regress on the order of hierarchical complexity of the items. Figure 4 shows the linear regression. As can be seen, the Order of Hierarchical Complexity of an item predicted the mean performance on that item with an R^2 of 0.97 (see Figure 16).

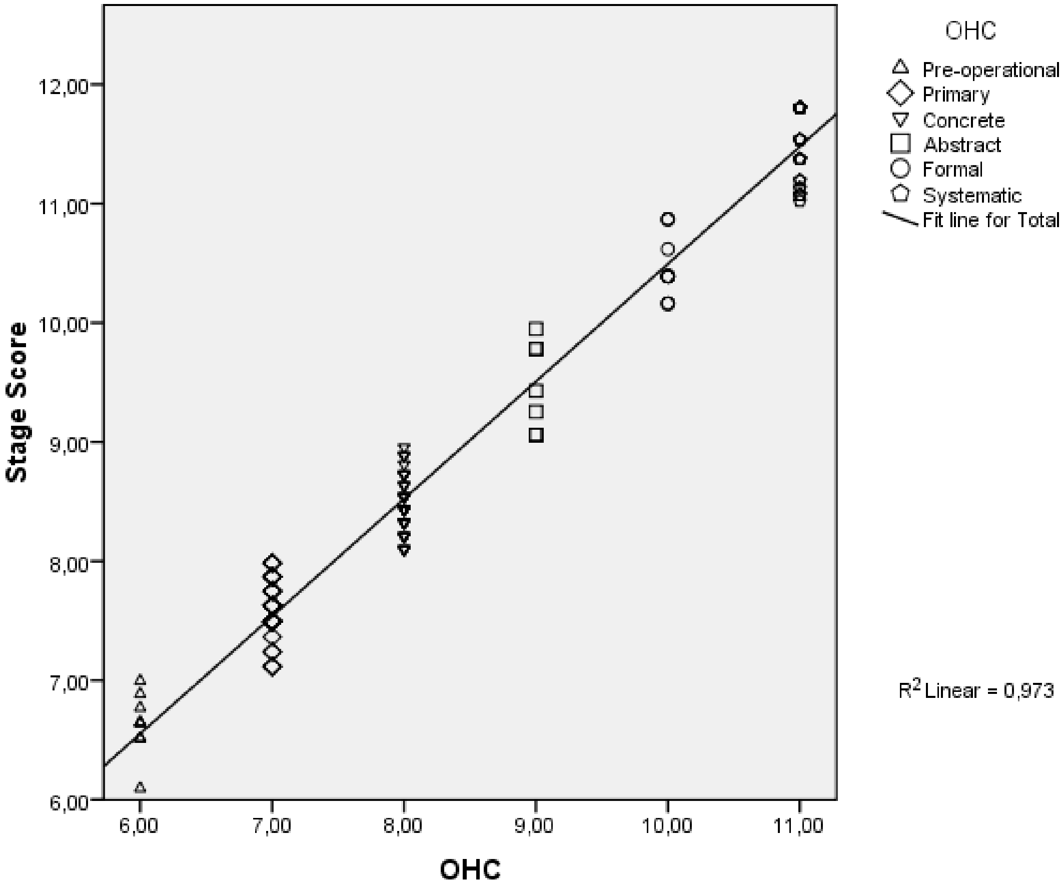


Fig. 16 Regression of Stage Scores on Order of Hierarchical Complexity

Conclusion

This study adds a new group of instruments with extremely high *r*'s between the order of hierarchical complexity used to predict the difficulty and the obtained difficulty. The difference between study 1 and 2 also shows the psychometric usefulness of constructing items with low horizontal complexity (number of actions) when what one is interested in is hierarchical complexity. Also of great import, is that these instruments test all the way down to the preoperational stage and go up through the systematic stage. It would be easy to make a metasystematic version by asking people to compare the degree of similarity between systems from the systematic order -- dissimilar, similar. Future studies should include higher stages.

The study also extends the application of the MHC and Skill Theory to another domain.

Table 4
Description of the IRDT demands by OHC

OHC	Name	What they do	How they do	
6	preoperational	Make very simple logical inductions, from single stimulus.	Proceeds from the identification and analysis of a group of single (equal) letters to a conclusion about an individual letter.	Distinguish single categories from each other (e.g. equal letters vs. different letter) in order to make a logical conclusion.
7	primary	Simple logical induction, from coordinated stimulus.	Proceeds from the identification of the relation between two coordinated letters, to a conclusion about a specific coordinated pair of letters.	Maps relations between pair of stimuli, and compare a series of paired relations in order to make a logical conclusion.
8	concrete	Logical induction from a system of mapped stimulus.	Proceeds from the analysis of *X* pair of coordinated letters, forming a system of relations within a single option, to a conclusion about a specific coordination of *X* pair of letters.	Analyze a system of relations between stimuli, and compare the systems to make a logical conclusion.

9	abstract	Logical induction carried out through the comparison of single abstract, general, class of systems.	Proceeds from the identification and comparison of variables out of finite classes, to a conclusion about a specific variable.	Distinguish single, general, abstract variables, in order to make a logical conclusion.
10	Formal	Logical induction from the coordinated abstract, general, class of systems.	Proceeds from the identification of the relation between two coordinated abstract variables, to a conclusion about a specific coordinated pair of variables.	Relationships are formed out of variables; mapping the relations to make a logical conclusion.
11	systematic	Logical induction from a system of mapped abstract, general, variables.	Proceeds from the analysis of *X* pair of coordinated abstract variables, forming a system of relations within a single option, to a conclusion about a specific coordination of *X* pair of abstract variables.	Analyze a system of relations between abstract, general variables, and compare the systems to make a logical conclusion.

References

Andrich, D. (1988). *Rasch models for measurement*. Sage series on quantitative applications in the Social Sciences, Beverly Hills.

Andrich, D. (2002). Understanding Rasch measurement: Understanding resistance to the data-model relationship in Rasch's paradigm: A reflection for the next generation. *Journal of Applied Measurement*, *3*(3), 325-359.

Andrich, D. (2004). Controversy and the Rasch model: a paradigm of incompatable paradigms. *Medical Care*, *42*(1).

Armon, C. (1984). Ideals of the good life and moral judgment: Ethical reasoning across the lifespan. In M. Commons & F. Richards & C. Armon (Eds.), *Beyond formal operations: Late adolescent and adult cognitive development, Vol 1*. (pp. 357-380). New York: Praeger.

Armon, C. & Dawson, T. L. (1997). Developmental trajectories in moral reasoning across the life-span. *Journal of Moral Education, 26,* 433-453.

Baillargeon, R. (1987). Object permanence in 3 1/2- and 4 1/2-month-old infants. *Developmental Psychology, 23*, 655-664.

Bart, W. (1971). The effect of interest on horizontal decalage at the stage of formal operations. *Journal of Psychology: Interdisciplinary and Applied*, *78*(2), 141-150.

Bernholt, S., Parchmann, I. & Commons, M. (March, 2008). *Hierarchical Complexity Applied to the Domain of Chemistry: An Educational Research and Modeling Approach*. Paper presented at the Society for Research in Adult Development, New York, NY.

Bidell, T. R., & Fischer, K. W. (1992). Beyond the stage debate: Action, structure, and variability in Piagetian theory and research. In R. J. Sternberg, C. A. Berg, R. J. Sternberg, C. A. Berg (Eds.) , *Intellectual development* (pp. 100-140). New York, NY US: Cambridge University Press.

Bock, R.D., & Jones, L.V. (1968). *The Measurement and Prediction of Judgment and Choice*. San Francisco: Holden Day.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.)*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Bowman, A. K. (1996a). *The relationship between organizational work practices and employee performance: Through the lens of adult development*. Unpublished doctoral dissertation. The Fielding Institute, Santa Barbara, CA.

Bowman, A. K. (1996b). Examples of task and relationship 4b, 5a, 5b statements for task performance, atmosphere, and preferred atmosphere. In M. L. Commons, E. A. Goodheart, T. L. Dawson, P. M. Miller, & D. L. Danaher, (Eds.) *The general stage scoring system (GSSS).* Presented at the Society for Research in Adult Development, Amherst, MA.

Broughton, J.M. (1984). Not beyond formal operations, but beyond Piaget. In M. Commons, F.A. Richards, and C. Armon (Eds.), *Beyond formal operations: Late adolescent and adult cognitive development,* Vol 1, (pp. 395-411). New York: Praeger.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytical studies*. New York: Cambridge University Press.

Chapman, M., & Lindenberger, U. (1988). Functions, operations, and decalage in the development of transitivity. *Developmental Psychology*, *24*, 542-551.

Chazan, S. (1972). Horizontal decalage in the concept of object permanence as a correlate of dimensions of maternal care. *Dissertation Abstracts International*, 32.

Colby, A., & Kohlberg, L. (1987a). *The measurement of moral judgment, Vol. 1: Theoretical foundations and research validation*. New York: Cambridge University Press.

Colby, A., & Kohlberg, L. (1987b). *The measurement of moral judgment, Vol. 2: Standard issue scoring manual*. New York: Cambridge University Press.

Commons, M. L. (2008). Introduction to the model of hierarchical complexity and its relationship to postformal action . *World Futures*, *64*, 305-320.

Commons, M. L., Krause, S. R., Fayer, G. A., & Meaney, M. (1993). Atmosphere and stage development in the workplace. In J. Demick & P. M. Miller (Eds.). *Development in the workplace* (pp. 199-220). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Commons, M. L., Lee, P., Gutheil, T. G., Goldman, M., Rubin, E. & Appelbaum, P. S. (1995). Moral stage of reasoning and the misperceived "duty" to report past crimes (misprision). *International Journal of Law and Psychiatry, 18*(4), 415-424.

Commons, M. L., Pekker, A. (2008). Presenting the formal theory of hierarchical complexity. *World Futures*, *64*, 375-382.

Commons, M. L., Rodriguez, J. A. (1990). “Equal access" without "establishing" religion: The necessity for assessing social perspective-taking skills and institutional atmosphere. *Developmental Review*, *10*, 323-340.

Commons, M. L., Rodriguez, J. A. (1993). The development of hierarchically complex equivalence classes. *Psychological Record, 43*, 667-697.

Commons, M. L., Rodriguez, J. A., Adams, K. M., Goodheart, E. A., Gutheil, T. G., & Cyr, E. D. (2006). Informed Consent: Do You Know It When You See It? Evaluating the Adequacy of Patient Consent and the Value of a Lawsuit. *Psychiatric Annals*, *36*, 430-435.

Commons, M. L., & Richards, F. A. (1984a). Applying the general stage model. In M. L. Commons, F. A. Richards, & C. Armon (Eds.), *Beyond formal operations. Late adolescent and adult cognitive development: Late adolescent and adult cognitive development,* Vol 1. (pp. 141-157). NY: Praeger.

Commons, M. L., Richards, F. A., & Kuhn, D. (1982). Systematic and metasystematic reasoning: A case for a level of reasoning beyond Piaget's formal operations. *Child Development, 53*, 1058-1069.

Commons, M., Goodheart, E., Pekker, A., Dawson, T., Draney, K., & Adams, K. (2008). Using Rasch scaled stage scores to validate orders of hierarchical complexity of balance beam task sequences. *Journal of Applied Measurement*, 9(2), 182-199.

Commons, M., Trudeau, E., Stein, S., Richards, F., & Krause, S. R. (1998). Hierarchical complexity of tasks shows the existence of developmental stages. *Developmental Review*, 18(3), 237-278.

Cook-Greuter, S. R. (1990). Maps for living: Ego-development theory from symbiosis to conscious universal embeddedness. In M. L. Commons, J. D. Sinnott, F. A. Richards, & C. Armon (Eds.). *Adult Development: Vol. 2, Comparisons and applications of adolescent and adult developmental models* (pp. 79-104). New York: Praeger.

Dawson, T. L. (2000). Moral reasoning and evaluative reasoning about the good life. *Journal of Applied Measurement, 1*, 372-397.

Dawson, T. L. (2001). Layers of structure: A comparison of two approaches to developmental assessment. *Genetic Epistemologist, 29* (4)*,* 1-10.

Dawson, T. L. (2002). New tools, new insights: Kohlberg's moral reasoning stages revisited. *International Journal of Behavioral Development, 26,* 154-166.

Dawson, T. L. (2003). A stage is a stage is a stage: A direct comparison of two scoring systems. *Journal of Genetic Psychology, 164*, 335-364.

Dawson, T. L. (2003). A stage is a stage is a stage: A direct comparison of two scoring systems. *Journal of Genetic Psychology, 164*, 335-364.

Dawson, T. L. (2004). Assessing intellectual development: Three approaches, one sequence. *Journal of Adult Development, 11,* 71-85.

Dawson, T. L. (2006). Stage-like patterns in the development of conceptions of energy. In X. Liu & W. Boone (Eds.), *Applications of Rasch measurement in science education* (pp. 111-136). Maple Grove, MN: JAM Press.

Dawson, T. L., & Wilson, M. (2004). The LAAS: A computerized developmental scoring system for small- and large-scale assessments. *Educational Assessment, 9*, 153-191.

Dawson, T. L., Xie, Y., & Wilson, M. (2003). Domain-general and domain-specific developmental assessments: Do they measure the same thing? *Cognitive Development*, *18,* 61-78.

Dawson, T., Goodheart, E., Draney, K., Wilson, M., & Commons, M. (2010). Concrete, abstract, formal, and systematic operations as observed in a 'Piagetian' balance-beam task series. *Journal of Applied Measurement*, 11(1), 11-23.

Dawson-Tunik, T. L. (2004). "A good education is…" The development of evaluative thought across the life-span. *Genetic, Social, and General Psychology Monographs, 130,* 4 112.

Dawson-Tunik, T. L., Commons, M., Wilson, M., & Fischer, K. (2005). The shape of development. *The European Journal of Developmental Psychology, 2,* 163-196.

Demetriou, A., & Kyriakides, L. (2006). The functional and developmental organization of cognitive developmental sequences. *British Journal of Educational Psychology*, 76(2), 209 242.

Demetriou, A., Christou, C., Spanoudis, G., & Platsidou, M. (2002). The development of mental processing: Efficiency, working memory, and thinking. *Monographs of the Society of Research in Child Development*, *67*, Serial Number 268.

Demetriou, A., Efklides, A., Papadaki, M., Papantoniou,G., & Economou, A. (1993). Structure and development of causal experimental thought: From early adolescence to youth. *Developmental Psychology, 29*, 480-497.

Demetriou, A., Mouyi, A., & Spanoudis, G. (2008). Modelling the structure and development of g. *Intelligence*, *36*(5), 437-454.

Embretson, S.E. and Reise, S. P. (2000). *Item response theory for psychologists.* London: Erlbaum.

Feldman, D.H. (2004). Piaget's stages: The unfinished symphony of cognitive development *New Ideas in Psychology*, *22*, 175-231.

Fischer, K. W. (2008). Dynamic cycles of cognitive and brain development: Measuring growth in mind, brain, and education. In A. M. Battro, K. W. Fischer, P. J. Léna, A. M. Battro, K. W. Fischer, P. J. Léna (Eds.), *The educated brain: Essays in neuroeducation* (pp. 127-150). New York, NY US:Cambridge University Press.

Fischer, K. W., & Bidell, T. R. (1998). Dynamic development of psychological structures in action and thought. In W. Damon, R. M. Lerner, W. Damon, R. M. Lerner (Eds.), *Handbook of child psychology: Volume 1: Theorectical models of human development (5th ed.)* (pp. 467-561). Hoboken, NJ US: John Wiley & Sons Inc.

Fischer, K. W., & Bidell, T. R. (2006). Dynamic development of action, thought, and emotion. In W. Damon & R. M. Lerner (Eds.), *Theoretical models of human development. Handbook of child psychology* (6th ed., Vol. 1, pp. 313-399). New York: Wiley.

Fischer, K. W., & Yan, Z. (2002a). The development of dynamic skill theory. In R. Lickliter & D. Lewkowicz (Eds.), *Conceptions of development: Lessons from the laboratory*. Hove, U.K.: Psychology Press.

Fischer, K. W., & Yan, Z. (2002b). Darwin's construction of the theory of evolution: Microdevelopment of explanations of variation and change of species. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition Processes in Development and Learning*. Cambridge, U.K.: Cambridge University Press.

Fischer, K.W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87*, 477-531.

Fischer, K.W. (1987). Relations between brain and cognitive development. *Child Development, 58*, 623-632.

Fischer, K.W., & Dawson, T. L. (2002). A new kind of developmental science: Using models to integrate theory and research. *Monographs of the Society for Research in Child Development*, *67* (1), 156-167.

Fischer, K.W., & Rose, S.P. (1994). Dynamic development of coordination of components in brain and behavior: A framework for theory and research. In G. Dawson & K.W. Fischer (Eds.). *Human behavior and the developing brain* (pp. 3-66). New York: Guilford Press.

Fischer, K.W., & Rose, S.P. (1999). Rulers, models, and nonlinear dynamics: measurement and method in developmental research. In G. Savelsbergh, H. van der Maas, and P. van Geert (Eds.), *Nonlinear developmental processes* (pp. 197-212).

Fischer, K.W., & Silvern, L. (1985). Stages and individual differences in cognitive development. *Annual Review of Psychology, 36*, 613-648.

Fischer, K.W., Hand, H.H., & Russell, S. (1984). The development of abstractions in adolescence and adulthood. In M. Commons, F.A. Richards, and C. Armon (Eds.), *Beyond formal operations: Late adolescent and adult cognitive development,* Vol 1, (pp. 43-73). New York: Praeger.

Fischer, K.W., Kenny, S.L., & Pipp, S.L. (1990). How cognitive processes and environmental conditions organize discontinuities in the development of abstractions. In C.N. Alexander, E.J. Langer, & R.M. Oetzel (Eds.), *Higher stages of development*. New York: Oxford University Press. Pp. 162-187.

Flavell, John H. (1963). *The Developmental Psychology of Jean Piaget*. Princeton, NJ: Van Nostrand.

Glas, C.A. (2007). *Multivariate and Mixture Distribution Rasch Models*. New York: Springer-Verlag.

Gomes, C. M. A. (2010). Estrutura fatorial da Bateria de Fatores Cognitivos de Alta-ordem (BAFACALO). *Avaliação Psicológica, 9*, 449-459.

Gomes, C.M.A. & Golino, H.F. (2009). Estudo exploratório sobre o Teste de Desenvolvimento do Raciocinio Indutivo (TDRI). In D. Colinvaux. *Anais do VII Congresso Brasileiro de Psicologia do Desenvolvimento: Desenvolvimento e Direitos Humananos.* (pp. 77-79). Rio de Janeiro: UERJ. Available in http://www.abpd.psc.br/files/congressosAnteriores/AnaisVIICBPD.pdf

Gomes, C. M. A. & BORGES, O. N. (2009). Qualidades Psicométricas do Conjunto de Testes de Inteligência Fluida. *Avaliação Psicológica*, *8*, 17-32.

Goodheart, E. A., Dawson, T. L. (June 1996). "A Rasch Analysis of Developmental Data from The Laundry Problem Task Series." Poster presented at the 11th Annual Adult Development Symposium, Boston, MA.

Goodheart, E. A., Dawson, T. L., Draney, K., Commons, M. L. (March 1997). "A Saltus Analysis of Developmental Data from The Laundry Problem Task Series." Poster presented at IOMW9, Chicago, IL.

Halford, G.S. (1989). Reflexions on 25 years of Piagetian cognitive developmental psychology, 1963 – 1988. *Human Development, 32,* 325-357.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(2), 38-47.

Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care*, 38(Suppl9), II60-II65.

Hartelman, P. A., van der Maas, H. J., & Molenaar, P. M. (1998). Detecting and modeling developmental transitions. *British Journal of Developmental Psychology*, 16(Pt 1), 97-122.

Hibbard, J., Collins, P., Mahoney, E., & Baker, L. (2010). The development and testing of a measure assessing clinician beliefs about patient self-management. *Health Expectations: An International Journal of Public Participation in Health Care & Health Policy*, *13*(1), 65 72.

Jackson, E., Campos, J.J. & Fischer, K.W. (1978). The question of decalage between object permanence and person permanence. *Developmental Psychology, 14.* 1-10.

Jamison, W. (1977). Developmental inter-relationships among concrete operational tasks: An investigation of Piaget's stage concept. *Journal of Experimental Child Psychology*, *24*(2), 235-253.

Joaquim, C. J. (2011). Developmental Stage of Performance in Reasoning About Bullying in School Age Youth. Doctoral dissertation, Nova Southeastern University.

Kallio, E., & Helkaman, K (1991). Formal operations and postformal reasoning: A replication. *Scandinavian Journal of Psychology*, 32, 1, 18-21.

Kitchener, K. S. & Fischer, K. W. (1990). A skill approach to the development of reflective thinking. In D. Kuhn (Ed.), *Developmental perspectives on teaching and learning thinking skills*. *Contributions to Human Development*: Vol. 21 (pp. 48-62).

Kitchener, K. S., & King, P. M. (1990). Reflective judgement: Ten years of research. In M. L. Commons, C. Armon, L. Kohlberg, F. A. Richards, T. A. Grotzer, & J. D. Sinnott (Eds.), *Beyond formal operations: Vol. 2. Models and methods in the study of adolescent and adult thought* (pp. 63-78). New York: Praeger.

Krantz, D.H., Luce, R.D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Vol. I: Additive and polynomial representations*. New York: Academic Press.

Kreitler, S., & Kreitler, H. (1989). Horizontal Decalage: A problem and its solution. *Cognitive Development, 4,* 89-119.

Lautrey, J., de Ribaupierre, A., & Rieben, L. (1985). Intraindividual variability on the development of concrete operations: Relations between logical and infralogical operations. *Genetic, Social, and General Psychology Monographs*, *111*(2), 167-192.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 32*(2), 103-122.

Linacre J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16* (2), 878

Linacre J. M. (2011). WINSTEPS. Rasch measurement computer program, Winsteps.com, Chicago.

Lourenço, O. (1998). Além de Piaget? Sim, mas Primeiro Além da Sua Interpretação Padrão! *Análise Psicológica*, *16*(*4*), p.521-552.

Lovell, C. W. (2002). Development and disequilibration: Predicting counselor trainee gain and loss scores on the *Supervisee Levels Questionnaire. Journal of Adult Development, 9*(3), 235-240.

Luce, R.D. & Tukey, J.W. (1964). Simultaneous conjoint measurement: a new scale type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1–27.

Marshall, P. E. (2009). *Positive psychology and constructivist developmental psychology: A theoretical enquiry into how a developmental stage conception might provide further insights into specific areas of positive psychology.* Unpublished Msc dissertation. University of East London, School of Psychology. Retrieved from http://devtestservice.org/about/articles.html

Martorano, S. (1977). A developmental analysis of performance on Piaget's formal operations tasks. *Developmental Psychology*, *13*(6), 666-672.

Miller, J. G., Bett, E. S., Ost, C. M., Commons, M. L., Day, J. M., Robinett, T. L., Ross, S. N., Marchand, H. & Lins, M. da Costa (June, 2008). *Finding the Relationships Among Moral Development Measures Using the Model of Hierarchical Complexity and Rasch Analysis*. Jean Piaget Society, Quebec City, Quebec, Canada.

Miller, J. G., Harrigan, W. J., Commons, M. L. & Commons-Miller, N. H. K. (November, 2008). *An Analysis of Causing Religious Belief and Atheism Instruments and Hierarchical Complexity*. Paper presented at the Association for Moral Education, Notre Dame University, South Bend, Indiana.

Miller, P. (2002). *Theories of developmental psychology (4th ed.)*. New York, NY US: Worth Publishers.

Miller, P. M., &. Lee, S. T. (June, 2000). *Stages and transitions in child and adult narratives about losses of attachment objects.* Paper presented at the Jean Piaget Society. Montreal, Québec, Canada.

Morra, S., Gobbo, C., Marini, Z., & Sheese, R. (2008). *Cognitive development: Neo Piagetian perspectives*. New York, NY: Taylor & Francis Group/Lawrence Erlbaum Associates.

Murray, F. (1969). Conservation in self and object. *Psychological Reports*, *25*(3), 941-942.

Murray, F., & Holm, J. (1982). The absence of lag in conservation of discontinuous and continuous materials. *Journal of Genetic Psychology*, *141*(2), 213-217.

Nummedal, S. (1971). The existence of the substance-weight-volume decalage. *Dissertation Abstracts International*, 31.

Panayides, P., Robinson, C., & Tymms, P. (2010). The assessment revolution that has passed England by: Rasch measurement, *British Educational Research Journal*, *36*(4), 611 626.

Pascual-Leone, J. (1984). Attentional, Dialectic, and Mental Effort: Toward an Organismic Theory of Life Stages, In Michael L. Commons, Francis A. Richards and Cheryl Armon (Eds.), *Beyond formal operations. Late adolescent and adult cognitive development,* Vol 1. (pp. 182-215). New York: Praeger.

Peeters, M.J. & Stone, G.E. (2009). An Instrument to Objectively Measure Pharmacist Professionalism as an Outcome: A Pilot Study. *The Canadian Journal of Hospital Pharmacy, 62*(3), 209-216.

Perry, W. G. (1970). *Forms of intellectual and ethical development in the college years*. New York: Holt, Rinehart, & Winston.

Rasch, G. (1960/1993). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research). Expanded edition (1980) with foreword and afterword by B.D. Wright, (1980).Chicago: MESA Press.

Richardson, A. M. & Commons, M. L. (July, 2008). *Accounting for Stage of Development on Mathematical and Physical Science Tasks*. Paper presented at the Society for Mathematical Psychology, Washington D.C.

Rijmen, F., De Boeck, P., & Van der Mass, H. J. (2005). An IRT Model with a Parameter Driven Process for Change. *Psychometrika*, *70*(4), 651-699.

Roberge, J. (1976). Developmental analyses of two formal operational structures: Combinatorial thinking and conditional reasoning. *Developmental Psychology*, *12*(6), 563 564.

Rose, S. P., & Fischer, K. W. (1998). Models and rulers in dynamical development. *British Journal of Developmental Psychology*, *16*(1), 123-131.

Salzberger, T. (2011). 'The role of the unit in physics and psychometrics' by Stephen Humphry—One small step for the Rasch model, but possibly one giant leap for measurement in the social sciences. Measurement: Interdisciplinary Research and Perspectives, 9(1), 59-61.

Scardamalia, M. (1977). Information processing capacity and the problem of horizontal Decalage : A demonstration using combinatorial reasoning tasks. *Child Development*, *48*(1), 28-37.

Schwartz, M. S., & Fischer, K. W. (2005). Building general knowledge and skill: Cognition and microdevelopment in science learning. In A. Demetriou & A. Raftopoulos (Eds.), *Cognitive developmental change: Theories, models, and measurement*. Cambridge, U.K.: Cambridge University Press.

Siegler, R., & Crowley, K. (1991). The gospel of Jean Piaget, according to John Flavell. *PsycCRITIQUES*, *36*(10), 829-831.

Siegler, R.S, (1981). Developmental sequences within and between concepts. *Monograph of The Society for Research in Child Development*, *46*(2), pp. 1-84.

Smith, L. (2002). From epistemology to psychology in the development of knowledge. In T. Brown, L. Smith, T. Brown, L. Smith (Eds.), *Reductionism and the development of knowledge* (pp. 201-228). Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Sonnert, G., & Commons, M. L. (1994). Society and the highest stages of moral development. *Politics and the Individual, 4*(1), 31-55.

Stein, Z., & Hiekkinen, K. (2009). Metrics, models, and measurement in developmental psychology. *Integral Review, 5*(1), 4-24.

Stein, Z., Dawson, T., & Fischer, K. W. (2010). Redesigning testing: Operationalizing the new science of learning. In M. S. Khine & I. M. Saleh (Eds.), *New Science of Learning: Cognition, Computers and Collaboration in Education* (pp. 207–224). New York: Springer.

Van der Maas, H. L., & Molenaar, P. C. M. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review,99,* 395–417.

Van Geert, P. & Steenbeek, H. (2005). Explaining after by before: Basic aspects of a dynamic systems approach to the study of development. *Developmental Review*, 25, 408 442.

Webb, R. (1974). Concrete and formal operations in very bright 6- to 11-year-olds. *Human Development*, *17*(4), 292-300.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wright, B.D. and Stone, M.H. (1979) The measurement model. Best Test Design: Rasch Measurement (pp. 1-17), Mesa Press: Chicago.

Yan, Z., & Fischer, K. W. (2007). Pattern emergence and pattern transition in microdevelopmental variation: Evidence of complex dynamics of developmental processes. *Journal of Developmental Processes*, *2*(2), 39-62.

2.2 Artigo 2

Identifying Developmental Stages transversally: Validity evidences of the Inductive Reasoning Developmental Test.

ABSTRACT

The current study investigates the structural validity of the Inductive Reasoning Developmental Test (IRDT) 3rd version, a fifty-six items test based on the Model of Hierarchical Complexity. The goal of the present paper is to check for developmental stages of reasoning. Three quantitative methodologies will be applied, each one covering a different aspect of the test structure: 1) Confirmatory Factor Analysis (CFA) will help reveal if items constructed to identify different stages form different latent variables, as predicted by the MHC, as well as check for second order unidimensionality; 2) Dichotomous Rasch Model will help reveal if the pattern of item difficulties form clusters separated by gaps; 3) A latent class model will help reveal how many discrete latent classes explain the distribution of item difficulties. The sample is composed by 1,459 Brazilian people (52.5% women, 47.5% men) aged between 5 to 86 years (M = 15.75, SD = 12.21). The results show a good fit to the Rasch Model (Infit mean = .96; SD = .17) with a high reliability estimate for items (1.00) and moderately high for people (.82). The item's difficulty distribution formed a clear seven cluster structure with gaps between them, presenting statistically significant differences in the 95% confidence interval level, as verified through one-sample t-test. The CFA showed an adequate data fit for a two-level model, being seven first-order factors and one second-order general factor [$\chi2$ (61) = 8832.594, p = .000, CFI = .96, RMSEA = .059]. The latent class analysis showed that the best model is the one with seven latent classes (AIC: 263.380; BIC: 303.887; Loglik: -111.690). These findings support the idea that the IRDT identifies seven developmental stages.

Keywords: Stages, Assessment, Validation, Development, Model of Hierarchical Complexity, Inductive Reasoning.

INTRODUCTION

Some authors have pointed the urge for the construction of metrics in developmental psychology (Fischer & Rose, 1999; Rose & Fischer, 1998; Van Geert & Steenbeek, 2005), with reliable, valid and accurate measures (Fischer & Dawson, 2002; Stein & Heikkinen, 2009). The post-piagetian researchers have been tackling this issue by developing and applying new methodologies, as well as creating innovative instruments that makes possible to reveal stage-like development (Commons, Goodheart, Pekker, Dawson, Draney & Adams, 2008; Bond & Fox, 2001; Dawson, 2000; Dawson-Tunik, Commons, Wilson, & Fischer, 2005; Dawson-Tunik, Goodheart, Draney, Wilson & Commons, 2010; Demetriou, Efklides, Papadaki, Papantoniou, & Economou,1993; Demetriou & Kyriakides, 2006;). Table 1 shows some studies that have been focusing in the empirical verification of developmental stages.

Although there's still a struggle whether development is continuous or discontinuous (stage-like), Fischer and his colleagues presented evidences for both kinds of developmental patterns (Fischer, Kenny, & Pipp, 1990; Fischer & Silvern, 1985; Fischer & Yan, 2002a,b; Schwartz & Fischer, 2005; Yan & Fischer, 2007). Continuous development relates to the sequence of steps needed in the construction of skills (i.e. microdevelopment), while discontinuity relates to abrupt, stage-like changes that marks the emergence of radically new kinds of control units of behavior and cognition (Fischer, 1980; Fischer & Rose, 1994; Fischer & Bidell, 1998, 2006; Fischer & Yan, 2002a).

Discontinuity can be checked by constructing instruments that focus on the hierarchical complexity of items, i.e. the organization of information in the form of action in two or more coordinated subtasks, rather than horizontal complexity, i.e. the number informational bits they demand to successful task completion (Commons, 2008; Commons & Pekker, 2008; Commons, Gane-McCalla, Barker, & Li, in press). As pointed by Golino, Gomes, Commons and Miller (in press), grouping items with the same hierarchical complexity within stages, and designing items with increasing hierarchical complexity between stages enables the empirical verification of discontinuity. The first strategy deals with item equivalence, which is important in order to avoid the elaboration of an anomalous scale that confuses its analysis (Fischer & Rose, 1999). The second strategy makes possible the identification of discontinuous development, with gaps between different orders of hierarchical complexity.

Instruments that do not control vertical complexity and do not focus on hierarchical complexity are less likely to adequately identify developmental stages.

Commons et al. (2008) and Dawson-Tunik et al. (2010), showed evidences of developmental stages of logical proportional reasoning using the Balance Beam task series, an instrument constructed following the Model of Hierarchical Complexity (citar), based on Piaget's balance beam task (Inhelder & Piaget, 1958). In the study published in 2008, they've employed the Dichotomous Rasch Model to verify if items constructed with the same hierarchical complexity would cluster their difficulties. Univariate statistics were applied, and the result showed that adjacent clusters presented statistically significant differences. In the study published in 2010 the authors verified the discontinuity of concrete, abstract, formal and systematic stage through the Saltus Model (Wilson, 1989). They've employed the Saltus Model because this is a logistic model (mixture extension of the Rasch Model) with a latent group parameter, and was constructed to determine "whether the difficulty of a group of items is significantly different for groups of persons who have different ability estimates" (Dawson-Tunik, Goodheart, Draney, Wilson, & Commons, 2010, p. 06). The result pointed to a two level model with gaps between the concrete/abstract and formal/systematic items. The lack of evidence to support a four level model indicates that the instrument needs revision, in order to identify what it intends to measure (four developmental stages with three gaps between them). The combination of Rasch Models and the Saltus model to verify discontinuities was also successfully used by Demetriou and Kyriakides (2006), as well as Bond and Fox (2001).

Golino, Gomes, Commons and Miller (in press) showed evidences of discontinuity by applying the dichotomous Rasch model on data collected through the Inductive Reasoning Developmental Test, a pencil-and-paper instrument also constructed based on the model of hierarchical complexity. The result showed six stages, distributed through the latent variable in six clusters of items difficulties, with significant gaps between them (verified through one sample t-test).

Table 1

Some studies investigating developmental stages

Reference	Instruments Used	N	Age (range, mean, standard-deviation)	Models Used	Domain	Reliability		Fit		Evidences of discontinuity
						Person	Item (or stage)	Person	Item (or stage)	
Demetriou, Efklides, Papadaki, Papantoniou, & Economou(1993)	Combinatorial ability battery, Experimentation ability battery, Hypothesis-evidence handling battery and model construction.	260	12-17, 14.44, 1.34	Confirmatory Factor Analysis (check for dimensionality)/ Rasch - Rating scale model/ Saltus model (to verify second-order discontinuities)	Causal-experimental reasoning	0.79	0.99	NA	4 first order factors plus one second order general factor [χ 2 (30) = 39.868, p = .108, CFI = .992]	Saltus model (the result showed that the loglikelihood value of the Saltus model increased only 10.26 from the loglikelihood of the Rasch Model, thus the authors concluded the abilities they were investigating is continuos rather than discontinuous)
Müller, Sokol & Overton (1999)	Class reasoning tasks and propositional reasoning tasks	80	6-13, 12, 1.41	Dichotomous Rasch Model	Logical Reasoning	NA	NA	NA	Infit t (M = -0.2, SD = 1.2)	Variable Maps

Dawson (2000)	Good Education Interview, Moral Judgement Interview and Evaluative Reasoning Interview	209	5-86, NA, NA	Partial Credit Model	Good Education Concepts, Moral Reasoning and Evaluative Reasoning	NA	0.98	NA	All items presented t < 2.0, and infit mean square ranging from 0.57 to 1.07.	Variable Maps
				Random coefficients multinomial logit model (RCML)	Good Education Concepts, Moral Reasoning and Evaluative Reasoning	NA	1.00	NA	All items presented t < 2.0, and infit mean square ranging from 0.61 to 1.48.	Variable Maps
Bond & Fox (2001)	Three Noelting tasks: Mixing Juices, Caskets task and coded orthogonal views	350	16 to adulthood, NA, NA	Dichotomous Rasch Model	Logical Reasoning/Visuo-spacial ability	NA	NA	NA	4 items out of 41 did not fit the model (infit and outfit t value exceeded the range between -2 and 2).	t tests between clusters of items/ variable maps (items difficulties)
Bond & Fox (2001)	Bond's logical operations test	150	Secundary students	Dichotomous Rasch Model	Logical Reasoning	0.81	0.94	infit mean square (M = 0.99, SD = 0.13)	infit mean square (M = 1.00, SD = 0.11)	Variable maps (items difficulties)
Bond & Fox (2001)	Piagetian Reasoning Task (PRTIII-Pendulum)	150	Secundary students	Dichotomous Rasch Model	Logical Reasoning	NA	NA	NA	infit mean square (M = 0.99, SD = 0.13)	Variable maps (items difficulties)
Bond & Fox (2001)	Mixing Juices Test	460	5-17, NA, NA	Polytomous Rasch Model/Saltus Model	Logical Reasoning	NA	NA	NA	NA	Variable maps (items difficulties)/Saltus Model

Dawson (2002)	Moral judgement Interviews scored using Kohlberg's Standard Issue Score System	996	5-86, 32, 16	Rasch - Partial Credit Model	Moral Reasoning	0.93	NA	12% of the sample exceeded the adequate infit t range (between - 2 and 2)	Infit Mean Square (M = 0.93, SD = 0.07)	Variable Maps (items' difficulties); 95% confidence intervals for each of the stage-item difficulty estimates were calculated from the standard errors.
Dawson, Xie & Wilson (2003)	Kohlberg's Standard lssue Scoring System and Hierarchical Complexity Score System	378	6-86,	Unidimensional and multidimensional partial credit analysis (Rasch Family of models)	Moral Reasoning	NA	NA	NA	NA	Variable maps (items difficulties)
Dawson-Tunik (2004)	Hierarchical Complexity Score System applied to the Good Education Interview	246	5-86, 26,67, 20.56	Rasch - Rating scale model	Good Education Concepts	.94	NA	0.5% exceeding fit range adopted (between - 2 and 2 infit z scores)	All items presented infit z scores less than 2.	category characteristic curve
Dawson-Tunik, Commons, Wilson, & Fischer (2005)	Hierarchical Complexity Score System and Lectical Assessment System applied to interviews about moral judgment	747	5-86, 25.38, 15.93	Rasch - Rating scale model	Moral Reasoning	.97	NA	3% exceeding fit range adopted (between - 2 and 2 infit z scores)	All items presented infit z scores less than 2.	category characteristic curve

Demetriou and Kyriakides (2006)	The comprehensive test of cognitive development	629	12.1-18.3, 15.7 (median), NA	Rasch Model / Saltus Model	Intelligence	0.92	0.99	Infit Mean Square mean of 0.99, outfit mean square mean of 1.07	Infit Mean Square mean of 0.99, outfit mean square mean of 1.07	Cluster Analysis / Saltus Model
Commons, Goodheart, Pekker, Dawson, Draney and Adams (2008)	The Balance Beam Task Series	121	7-66, 29.22, 12.98	Dichotomous Rasch Model	Logical Reasoning	NA	0.98	NA	Infit Mean Square ranging from 0 .24 to 1.41 (M= 0.59)	Variable maps, univariate analysis of stage spacing.
Dawson-Tunik, Goodheart, Draney, Wilson and Commons (2010)	The Balance Beam Task Series	121	7-56, 29.2, 12.98	Dichotomous Rasch Model/Saltus Model	Logical Reasoning	0.77	0.97	Infit Mean Square mean of 0.95, standard deviation of 0.64.	Infit Mean Square mean of 0.94, standard deviation of 0.13.	Variable Map/Saltus Model
Golino, Gomes, Commons and Miller (in press)	The Inductive Reasoning Developmental Test (IRDT) 1st version (study 1)	167	6-58, 18.9, 9.7	Dichotomous Rasch Model	Logical Reasoning	0.95	0.97	NA	Infit Mean Square ranging from 0 .39 to 1.69 (M= 0.87, SD= 0.28)	Variable Map/One-Sample t-test, with 95% confidence interval comparing stage means
	The Inductive Reasoning Developmental Test (IRDT) 2nd version (study 2)	188	6-65, 21.45, 14.31	Dichotomous Rasch Model	Logical Reasoning	0.95	0.99	NA	Infit Mean Square ranging from 0 .56 to 1.46 (M= 0.94, SD= 0.22)	Variable Map/One-Sample t-test, with 95% confidence interval comparing stage means

As shown in table 1, the Rasch models have been vastly used in the post-Piagetian literature. Some studies present evidences of stages as clusters of items difficulties, while other adds a specific latent class model, the Saltus model, in order to strengthen the evidences of discontinuity. A third kind of study also applies the Rasch Models, not in tests or tasks, but in categories created from interviews thought score systems. Dawson (2000, 2002), Dawson, Xie and Wilson (2003), Dawson-Tunik (2004) and Dawson-Tunik, Commons, Wilson and Fischer (2005), employed the Hierarchical Complexity Score System in moral judgement interviews and showed its construct and congruent validity, internal consistency and inter-rater reliability. The above studies also showed that in spite of measuring the same latent variable, the domain-free scales (HCSS) present better internal consistency, allow meaningful comparisons across domains and contexts, and enable the examination of the relationship between developmental stages and conceptual content. The evidence of stages comprised the distribution of the response categories through the latent variable, being each category a specific stage of moral reasoning. The result shows that the same response categories (stage) are clustered together and present gaps between adjacent categories. As pointed by Golino, Gomes, Commons and Miller (in press), despite its importance in guiding research and practice, the application of the interview-and-score methodology demand various trained scoring analysts, with high agreement between them, require a considerable time for large scale assessment and are vulnerable to subjective bias. The construction of objective tests and tasks brings speed and lower cost-procedures for evaluating large samples. Adopting the Model of Hierarchical Complexity as a reference for item construction, controlling horizontal complexity within stages and increasing vertical complexity between stages (a process Golino, Gomes, Commons and Miller call *verticalization*), and applying quantitative methodologies that can help revealing discontinuities is testable way of constructing metrics in developmental psychology.

The goal of the present paper is to check for developmental stages of reasoning, studying the structural validity of the Inductive Reasoning Developmental Test (IRDT), 3rd version. Three quantitative methodologies will be applied, each one covering a different aspect of the test structure: 1) Confirmatory Factor Analysis will help reveal if items constructed to identify different stages form different latent variables, as predicted by the MHC, as well as check for second order unidimensionality; 2) Dichotomous Rasch Model will help reveal if the pattern of item difficulties form clusters separated by gaps; 3) A latent class model will help reveal how many discrete latent classes explain the distribution of item difficulties. Six predictions will be tested through the application of the three methodologies above:

1) Each group of eight items (one group for each stage) are very close to each other in terms of difficulty, so we can visually verify seven clusters of items with gaps between them, using the Wright Map (Rasch Model);
2) Each cluster of item's difficulties are significantly different from the next adjacent cluster of items (Rasch Model plus univariate statistics);
3) Each group of eight items are explained by a latent variable representing a specific stage, so seven latent variables will be found (Confirmatory Factor Analysis);
4) The seven latent variables are explained by a general second order latent variable (Confirmatory Factor Analysis);
5) One first-order general factor explaining the observable answers to the 56 items will not present adequate fit;
6) The item difficulties are explained by seven latent classes (Latent Class Model);

If we fail to visually identify seven clusters of item difficulties with gaps between them, through the application of the Rasch Model, then prediction 1 will be refuted, and the test will need revision. This revision will also be demanded if at least one item does not fit the Rasch Model, and/or if it falls in a cluster other than the one it was intended to measure. If we fail to identify statistically significant differences between the clusters of items' difficulties, then prediction 2 will be refuted. If we fail to identify seven latent variables, each one composed of eight items constructed to identify the same stage, then prediction 3 will be refuted. If we fail to identify a second-order general factor, prediction 4 will be rejected. However, if we identify a first-order general factor then prediction 5 will be refuted. Finally, if we fail to identify seven latent classes, then prediction 6 will also be rejected. Except for prediction 4, which is not a matter of stages but of unidimensionality, the greater the number of non-refuted predictions, the stronger the evidence supporting the existence of discontinuity, as assessed by the IRDT.

METHOD

Participants

The IRDT was administered to a convenience sample composed by 1,459 Brazilian people (52.5% women, 47.5% men) aged between 5 to 86 years (M = 15.75, SD = 12.21). The sample was intentionally broad, and had a distribution of 21.4% from 5 to 10 years old, 62.7% from 11 to 17 years old, 7.5% from 18 to 29, 6.4% from 30 to 59 and 2.1% older than 60 years old. All the participants were from the city of Belo Horizonte, state of Minas Gerais, Brazil.

Instrument

The Inductive Reasoning Developmental Test – IRDT (Gomes & Golino, 2009) is a pencil-and-paper instrument design to assess developmentally sequenced and hierarchically organized inductive reasoning. It is an extension, in terms of complexity, from the *Indução* test, which compose the fluid intelligence test kit (Gomes & Borges, 2009) of the Higher-Order Cognitive Factors Kit (Gomes, 2010). The domain of inductive reasoning was used because it is one of the best indicators of fluid intelligence (Carroll, 1993). The construction of the IRDT, from the original *Indução* items, is due to a larger challenge that concerns the construction of an intelligence battery to identify developmental stages.

The sequence of IRDT was constructed based on the MHC and on Fischer's Dynamic Skill Theory. Formally, the MHC stipulates that one task is more hierarchically complex than another task if all of the following are true.

a) It is defined in terms of two or more lower-order task actions. In mathematical terms, this is the same as a set being formed out of elements. This creates the hierarchy.

i. $A = \{a, b\}$, where *a* and *b* are "lower" than A and compose the set A;

ii. $A \neq \{A, ...\}$, where the A set cannot contain itself. This means that higher order tasks cannot be reduced to lower order ones. For example, postformal task actions cannot be reduced to formal ones.

b) It organizes lower order task actions. In mathematics' simplest terms, this is a relation on actions. The relations are order relations:

i. $A = (a, b) = \{a, \{b\}\}$ an ordered pair

c) This organization is non-arbitrary. This means that there is a match between the model that designates orders and the real world orders. This can be written as: Not P(a,b), not all permutations are allowed (see Commons & Pekker, 2008).

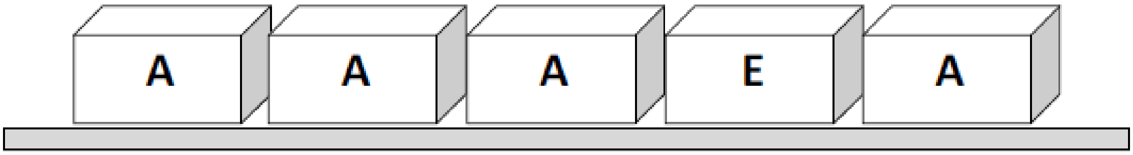
In sum, the MHC postulates that actions at a higher order of hierarchical complexity: 1) are defined in terms of two, or more, lower-order actions; 2) organize and transform those actions, not just combine them in a chain; and 3) produce organizations of lower-order actions that are new and not arbitrary.

The first two versions of the IRDT (Golino, Gomes, Commons & Miller, in press) was designed to identify six developmental stages (or levels), that will be named based in the MHC, respectively: Pre-operational; Primary; Concrete; Abstract; Formal; and Systematic. Each stage is composed of eight items with the same order of hierarchical complexity (OHC), for a total of forty-eight items. Each item is composed of four letters, or sequence of letters, with a specific rule (correct items), plus one letter or sequence with a different rule (exception). The task is to discover which letter or sequence is the exception. The 3rd version of the IRDT keeps its original idea, but implements two main changes: 1) reformulates the Abstract, Formal and Systematic items and 2) adds a new stage, namely Metassystematic. The changes from the previous version (Golino, Gomes, Commons, & Miller (in press) will be presented while describing the 3rd version, employed in the current study.

Pre-operational Items (Pre-op):

The eight Pre-op items demand the participants to make very simple logical inductions, from single stimulus. The participants need to proceeds from the identification and analysis of a group of single (equal) letters to a conclusion about an individual letter. In other words, they demand people to distinguish single categories from each other (e.g. equal letters vs. different letter) in order to make a logical conclusion (see Fig. 1).

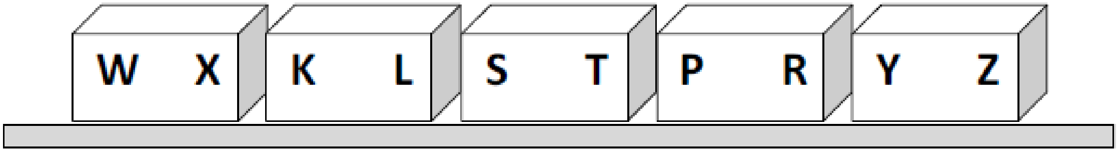
Fig. 1 Example: item 1, pre-operational stage



Primary Items (Prim):

The eight primary items demand the participants to make simple logical induction, from coordinated stimulus. The participants need to proceed from the identification of the relation between two coordinated letters, to a conclusion about a specific coordinated pair of letters. Mapping the relations between pair of stimuli, and comparing a series of paired relations in order to make a logical conclusion is demanded by the primary items (see Fig 2).

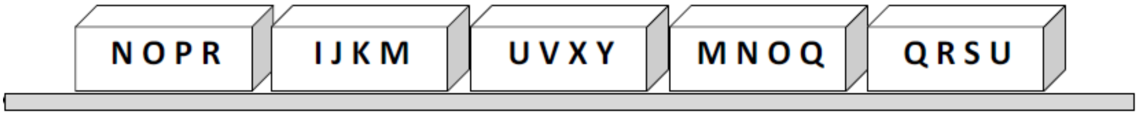
Fig. 2 Example: item 9, primary stage



Concrete items (conc):

The eight concrete items demand the participants to make a logical induction from a system of mapped stimulus. The participants will need to proceed from the analysis of X pair of coordinated letters, forming a system of relations within a single option, to a conclusion about a specific coordination of X pair of letters. Analyze a system of relations between stimuli, and compare the systems to make a logical conclusion, are demanded by the concrete items (see Fig 3).

Fig. 3 Example: item 17, concrete stage



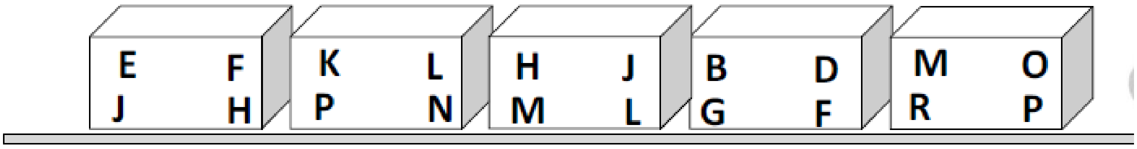
The previous version of the IRDT (Golino, Gomes, Commons, & Miller, in press) presented a table with codes (Greek letters), each one representing a coordination of two sets of four letters. The table was, then, the Abstract items. The formal items were just the coordination of two Greek letters, while the Systematic items were the coordination of two groups with two Greek letters. In sum, each higher stage was composed of a coordination of stimulus from previous stage.

In spite of the adequate fit to the Rasch model (Infit Mean-Square: M= .87; SD = .28), the 2nd version of the IRDT had the little issue of the Abstract table, in which the formal and systematic items relied on. This particular characteristic is problematic since it may create a local dependency of the formal and systematic items. This issue was solved in the 3rd version. We'll see how the abstract, formal, systematic and the new metassystematic items look like in the next paragraphs.

Abstract items (abs):

The eight abstract items demand the participants to make a logical induction through the comparison of single abstract, general, class of systems. The systems are composed of four letters displayed in a squared design (see Fig. 4).

Fig. 4 Example: item 25, abstract stage



The participant needs to verify how the letters are related to each other in a system, and compare different systems, choosing the one which does not follow the same pattern of the other four. However, differently from the previous (concrete) items, the system is closed, so the 1st and the last letter are also related. Let's take the example of figure 4. In the first option we have E, F, H and J displayed in a square design. The participant needs to analyze the relationship between E and F (no intermediary letter), F and H (one intermediary letter), H and J (one intermediary letter), as well as between J and E (four intermediary letters). The systems are: E-F-H (system 1), H-J-E (system 2), and they are reversible, so it goes forth and back (from E to J and J to E). This option shows an abstract pattern of relationship between the systems, as we can verify in the figure 5 below:

Fig. 5 Identifying relationships on item 25

System 1				System 2			
E	F	H		H	J	E	Abstract relationship between systems : 01-14
	0	1			1	4	
K	L	N		N	P	K	
	0	1			1	4	
H	J	L		L	M	H	Abstract relationship between systems : 11-04
	1	1			0	4	
B	D	F		F	G	B	
	1	1			0	4	
							Abstract relationship between systems : 10-14
M	O	P		P	R	M	
	1	0			1	4	

The participant must be able to verify the abstract pattern of relationship between the systems, as represented in the above figure by the number of intermediate letters between a pair of letters. Two patterns (single abstract, general, class of systems) appear in four options, and a third pattern appears in one option, the one that must be indicated in the answer (option MOP-PRM).

Formal items (form):

The eight formal items demand the participants to make a logical induction through the analysis of coordinated abstract, general, class of systems. The participants need to proceed from the identification of the relation between three coordinated abstract variables (see figure 6) representing one option, to a conclusion about one specific option that does not follow the pattern of the others.

Fig. 6 Example: item 33, formal stage

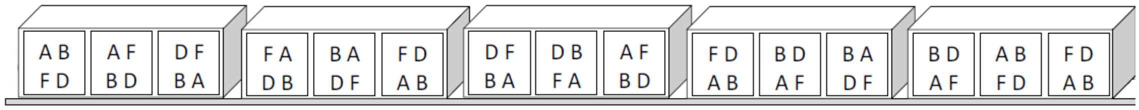
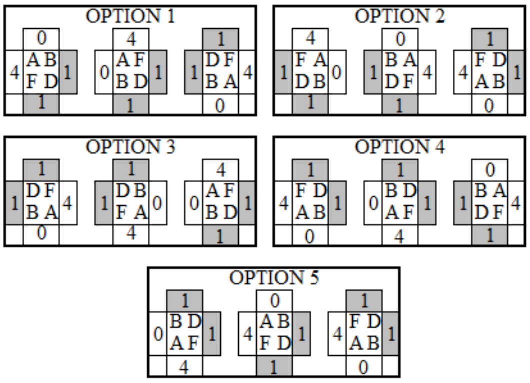


Figure 7 below exemplifies the item demand. Four options present a pattern where letters with distance 1 are at the same position in the first two abstract variables, and at the opposite position in the third abstract variable (options 1 to 4). The only option that does not follow this pattern is option 5.

Fig. 7 Identifying the relationships on item 33


OPTION 1
0 4 1
4 AB FD 1 0 AF BD 1 1 DF BA 4
1 1 0
OPTION 2
4 0 1
1 FA DB 0 1 BA DF 4 4 FD AB 1
1 1 0
OPTION 3
1 1 4
1 DF BA 4 1 DB FA 0 0 AF BD 1
0 4 1
OPTION 4
1 1 0
4 FD AB 1 0 BD AF 1 1 BA DF 4
0 4 1
OPTION 5
1 0 1
0 BD AF 1 4 AB FD 1 4 FD AB 1
4 1 0

Systematic items (sys):

The eight systematic items demand the participants to make a logical induction through the comparison of a system of mapped abstract, general, variables. The participants need to proceed from the analysis of 2 pair of coordinated abstract variables, forming a system of relations within a single option, to a conclusion about a specific coordination of 2 pair of abstract variables (see figure 8).

Fig. 8 Example: item 41, systematic stage

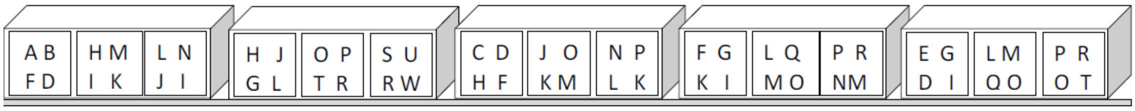
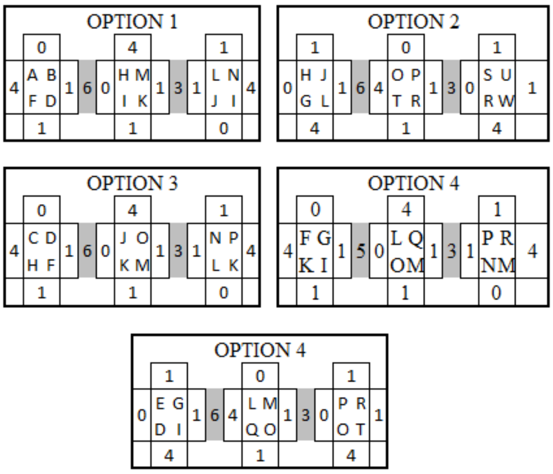


Figure 9 below exemplifies the item demand. Four options (1, 2, 3 and 5) present a pattern where the first pair of mapped abstract variables have distance 6, e.g. A to H, while the second pair have distance 3, e.g. H to L. The only option that does not follow this pattern is option 4.

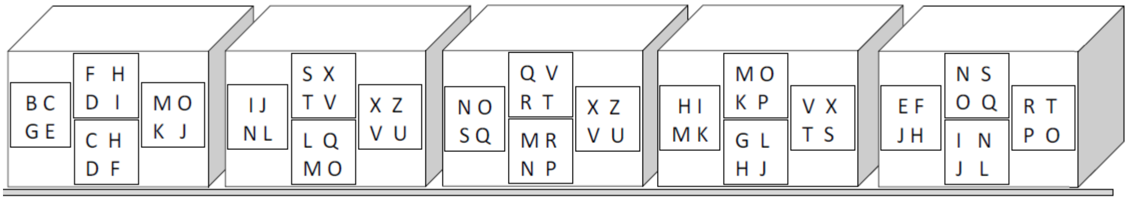
Fig. 9 Identifying relations on item 41


OPTION 1
0 4 1
4 AB FD 1 6 0 HM IK 1 3 1 LN JI 4
1 1 0
OPTION 2
1 0 1
0 HJ GL 1 6 4 OP TR 1 3 0 SU RW 1
4 1 4
OPTION 3
0 4 1
4 CD HF 1 6 0 JO KM 1 3 1 NP LK 4
1 1 0
OPTION 4
0 4 1
4 FG KI 1 5 0 LQ OM 1 3 1 PR NM 4
1 1 0
OPTION 4
1 0 1
0 EG DI 1 6 4 LM QO 1 3 0 PR OT 1
4 1 4

Metassystematic items (met)

The eight metassystematic items demand the participants to make a logical induction through the comparison of systems of abstract systems (figure 10). The participants need to identify the relations among systems, and figure out what's the similarity between them all. The first option in figure 10 shows B presenting distance 3 from F; F presenting distance 6 from M and -2 from C. Summing 3, 6 and -2, we have the broad rule of the systems, i.e. 7. All the other options present the same broad rule, except option 3, since E presents distance 2 from Q; Q presents distance 6 from X and -3 from M. Summing these distances we find 5 instead of 7.

Fig. 10 Example: item 49, metassystematic stage



In short, the metassystematic items coordinate two systematic components. By its turn, the systematic items coordinate two formal components, while the formal items coordinate two abstract components. The abstract items coordinate two concrete components. The concrete items coordinate two primary components. Finally, the primary items coordinate two pre-operational components.

DATA ANALYSIS

In order to verify if the six predictions presented in the introduction are true, we'll apply three different quantitative techniques: Confirmatory Factor Analysis, Dichotomous Rasch Model and Latent Class Model. Below will be briefly described the techniques and/or the procedures used to verify the data-fit to each specific model.

CFA:

The confirmatory factor analysis will be used through the software Mplus 5.2. Data fit to the hypothesized model (one first-order latent variables explaining each group of 8 items, in a total of seven first-order latent variables, and a second-order general factor explaining the seven first-order latent variables) as well as to the alternative model (a general first-order latent variable explaining the 56 items) will be verified using the root mean-square error of approximation (RMSEA) and the comparative fit index (CFI). A good data fit is indicated by a RMSEA shorter than .08, and a CFI equal to or greater than .90. The alternative model, with one first-order general factor, is not expected to fit the data.

Dichotomous Rasch Model

The Rasch Model will be applied using the software Winsteps (Linacre, 1999, 2011). Among its benefits, it makes possible to reduce all the items into a unique developmental scale (Demetriou & Kyriakides, 2006), collapsing at the same latent trait person's abilities and item's difficulty (Bond & Fox, 2001; Embreston & Reise, 2000; Glas, 2007), and also enables the verification of hierarchical sequences of both item and person, being relevant to stage identification (Dawson, Xie & Wilson, 2003). To verify the adjustment of the data to the model, the information-weighted fit mean-square statistic (infit) will be used. Values between .5 and 1.5 logits are considered productive for measurement (Wright & Linacre, 1994). The unidimensionality of the checked by a number of procedures, each one complementing the other (see Tennant & Pallant, 2006). Here, unidimensionality will be addressed using only the model fit statistics – i.e. if the data fit the model, one of the consequences is the linearity of the measure, its unidimensionality, and so on – and the principal contrast, which can be verified through the percentage of variance explained by measures, and by the percentage of unexplained variance in the first contrast. The former should be closer to or greater than 60% (Peeters & Stone, 2009), while the latter should be closer to or less than 10%.

In the second part of the analysis, the spacing of Rasch scores between items of adjacent clusters will be verified using one-sample t-test with a 95% confidence interval. The Rasch scores represent the difficulty of an item (δ), which is its location at the latent variable continuum. It would have been good to compare the Rasch Scores for every item from adjacent clusters, being each cluster composed of eight items with the same hierarchical complexity, but because there were so many items, this would have produced too many comparisons. To reduce the number of comparison pairs, each item's Rasch score was subtracted from the mean Rasch score of the items from the next higher order of complexity (cluster). This calculation is represented by the formula 1:

$$\overline{\mathrm{X}}_{k+1} - \delta_{i_k} = Adj\delta_{i_k} \tag{1}$$

where $\overline{\mathrm{X}}_{k+1}$ is the mean of the next higher order of complexity (or Stage k+1), and δ_{i_h} is the difficulty of item *i* from order *k* (or Stage k) , producing the adjusted difficulty of item *i.*

In order to verify if the one-sample t-test can be computed, Kolmogorov-Smirnov test will check the normality of the adjusted difficulty of items.

Latent Class Model

Latent Class Models (LCM), or finite mixture models, are a set of probabilistic models that specifies a finite number of *n* discrete unobservable variables that causes the observable outcomes. The outcomes are assumed to be independent conditional on the latent class (Visser & Speekenbrink, 2010). The LCM will be applied in our data using as "outcomes" the 56 IRDT's item difficulties as estimated by the Rasch Model. The IRDT was constructed to identify 7 developmental stages, and we've predicted that each group of eight items (one group for each stage) are very close to each other in terms of difficulty, forming seven clusters separated by gaps. So, the item's difficulty distribution is expected to be explained by seven latent classes.

In order to apply the LCM, the depmixS4 package (Visser & Speekenbrink, 2010) of the R software will be employed. According to the authors, "although depmixS4 was designed to deal with longitudinal or time series data, for say T >100, it can also handle the limit case when T = 1. In this case, there are no time dependencies between observed data and the model reduces to a finite mixture or latent class model" (Visser & Speekenbrink, 2010, p.2). Eight models will be estimated, from 1 to 8 latent classes. In order to choose the best model to our data, two indexes will be employed: Akaike's Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978).

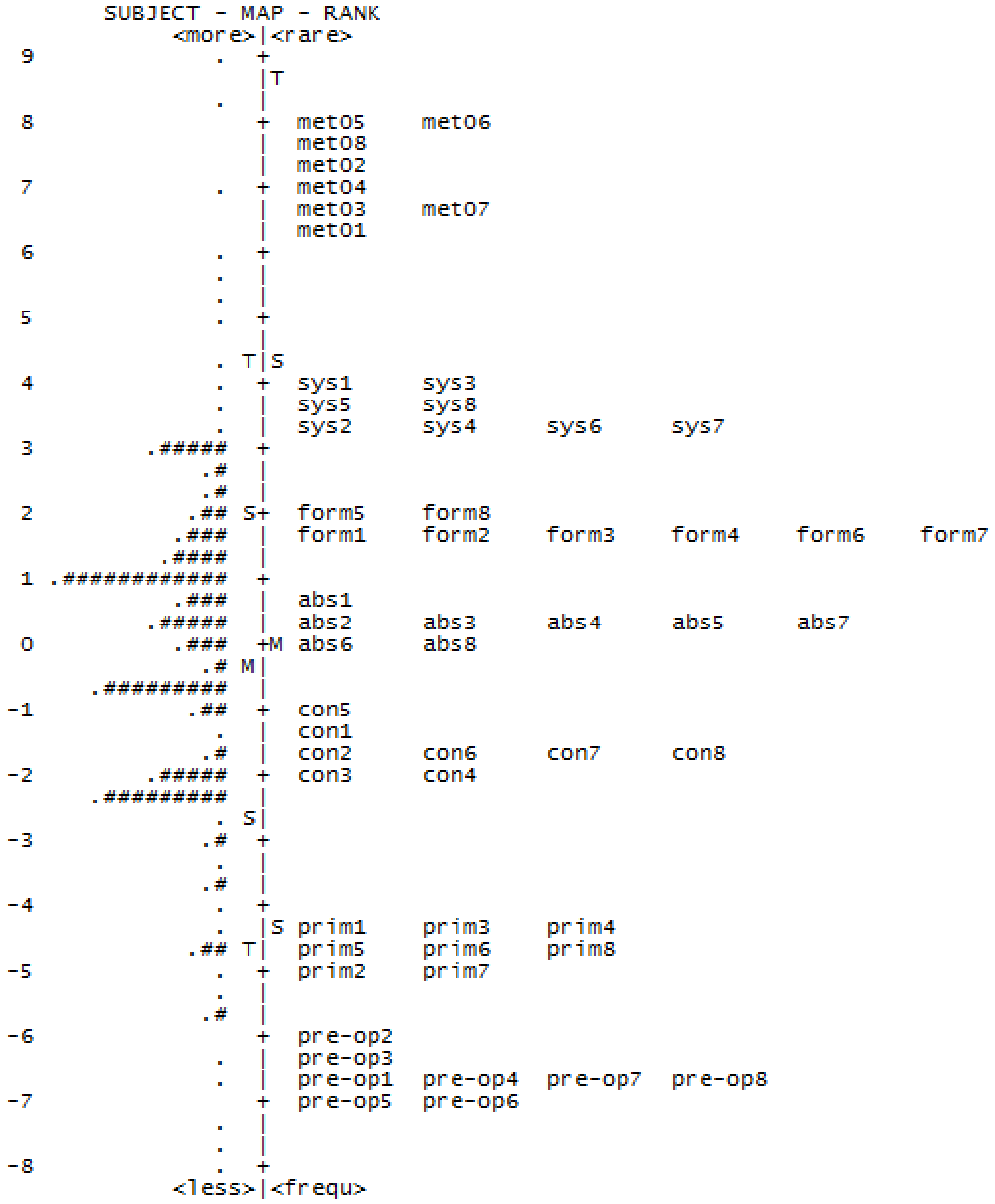
Since the depmixS4 package uses the expectation-maximization (EM) algorithm to maximize the log-likelihood, the AIC and BIC values can range due to random initialization of this algorithm. In this case, several iterations are necessary to estimate the global minimum of the AIC and BIC, instead of the local minimum (Haughton, Legrand, & Woolford, 2009). So, two hundred models will be used for estimating the AIC and BIC for each number of latent classes, from 1 to 8.

RESULTS

The CFA showed an adequate data fit for the two-level model, being seven first-order factors and one second-order general factor [χ^2 (61) = 8832.594, p = .000, CFI = .96, RMSEA = .059]. Factor weights on the first order factors varied from .66 to .99 (M = .90, SD = .08). The factor weights of latent variables on the second order general factor were .47 (pre-operational latent variable), .81 (primary latent variable), .78 (concrete latent variable), .77 (abstract latent variable), .62 (formal latent variable), .41 (systematic latent variable) and .017 (metassystematic latent variable).So, prediction 3 and 4 are cannot be refuted. Moreover, the first-order general factor model did not presented an adequate data-fit [χ^2 (61) = 8832.594, p = .000, CFI = .885, RMSEA = .105], also not refuting prediction 5.

The Rasch analysis showed a reliability of 1.00 for the 56 items, with an infit mean of .96 (SD = .17; Max = 1.32; Min = .72), falling within the acceptable fit range. The person reliability was .82, which is estimated to indicate the degree to which a person's response pattern conforms to the difficulty structure of the measure (Hibbard, Collins, Mahoney & Baker, 2009). The principal contrast showed that the raw variance explained by measures (modeled) is 70.3%, and that the unexplained variance in the first contrast (modeled) is 5.6%, suggesting that the instrument can be thought of as unidimensional.

Fig. 11 Wright Map



```
        SUBJECT - MAP - RANK
              <more>|<rare>
  9              .  +
                    |T
                 .  |
  8                 +  met05    met06
                    |  met08
                    |  met02
  7              .  +  met04
                    |  met03    met07
                    |  met01
  6              .  +
                 .  |
                 .  |
  5              .  +
                    |
                 . T|S
  4              .  +  sys1     sys3
                 .  |  sys5     sys8
                 .  |  sys2     sys4     sys6     sys7
  3         .#####  +
                .#  |
                .#  |
  2            .## S+  form5    form8
              .###  |  form1    form2    form3    form4    form6    form7
             .####  |
  1  .############  +
              .###  |  abs1
            .#####  |  abs2     abs3     abs4     abs5     abs7
  0           .###  +M abs6     abs8
                .# M|
        .#########  |
 -1            .##  +  con5
                 .  |  con1
                .#  |  con2     con6     con7     con8
 -2         .#####  +  con3     con4
        .#########  |
                 . S|
 -3             .#  +
                 .  |
                .#  |
 -4              .  +
                 .  |S prim1    prim3    prim4
               .## T|  prim5    prim6    prim8
 -5              .  +  prim2    prim7
                 .  |
                .#  |
 -6                 +  pre-op2
                 .  |  pre-op3
                 .  |  pre-op1  pre-op4  pre-op7  pre-op8
 -7                 +  pre-op5  pre-op6
                 .  |
                 .  |
 -8              .  +
              <less>|<frequ>
```

The Wright map (figure 11) illustrates the scale for the 56 items of the IRDT with item difficulties (on the right) and person measures (on the left) calibrated on the same scale. It is visually possible to identify seven item clusters with gaps between them. The adjusted item difficulty was computed accordingly to formula 1 presented in the methods section, and each group of eight adjusted scores presented normal distributions (see table 2).

Table 2

One-Sample Kolmogorov-Smirnov Test

		Preop/Prim	Prim/Conc	Conc/Abs	Abs/Form	Form/Syst	Syst/Meta
N		8	8	8	8	8	8
Normal Parameters[a,,b]	Mean	1,9988	2,9938	1,9300	1,4438	1,8325	3,6100
	Std. Deviation	,35504	,27313	,30458	,16379	,13169	,25444
Most Extreme Differences	Absolute	,270	,136	,263	,245	,204	,243
	Positive	,142	,136	,263	,245	,204	,126
	Negative	-,270	-,120	-,210	-,127	-,156	-,243
Kolmogorov-Smirnov Z		,763	,385	,744	,693	,577	,687
Asymp. Sig. (2-tailed)		,606	,998	,637	,723	,893	,733

The One-Sample t-test, with 95% confidence interval, shows that the comparisons of difficulty between Pre-operational and Primary, Primary and Concrete, Concrete and Abstract, Abstract and Formal, Formal and Systematic, as well as between Systematic and Metassystematic were significant (see table 3).

Table 3

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	T	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Preop/Prim	15,923	7	,000	1,99875	1,7019	2,2956
Prim/Conc	31,002	7	,000	2,99375	2,7654	3,2221
Conc/Abs	17,922	7	,000	1,93000	1,6754	2,1846
Abs/Form	24,931	7	,000	1,44375	1,3068	1,5807
Form/Syst	39,360	7	,000	1,83250	1,7224	1,9426
Syst/Meta	40,129	7	,000	3,61000	3,3973	3,8227

The result of the LCM shows that the best model is the one with seven latent classes (table 4), since it presented the lowest AIC and BIC value.

Table 4

Number of Latent Classes	Loglikelihood	AIC	BIC
1	-162.571	329.141	333.192
2	-156.442	322.844	332.970
3	-150.315	316.630	332.832
4	-143.919	309.838	332.110
5	-134.536	297.073	325.428
6	-120.986	275.973	310.404
7	-111.690	263.380	303.887
8	-109.909	265.819	312.402

Figure 12 and 13 show variability of AIC and BIC indexes by models generated. The two-class model have its AIC and BIC values ranging from 0 to 199, the three-class model from 200 to 399, the four-class model from 400 to 599, the five-class model from 600 to 799, the six-class model from 800 to 999, the seven-class model from 1000 to 1999, and the eight-class model from 1200 to 1399.

Fig. 12 AIC plot

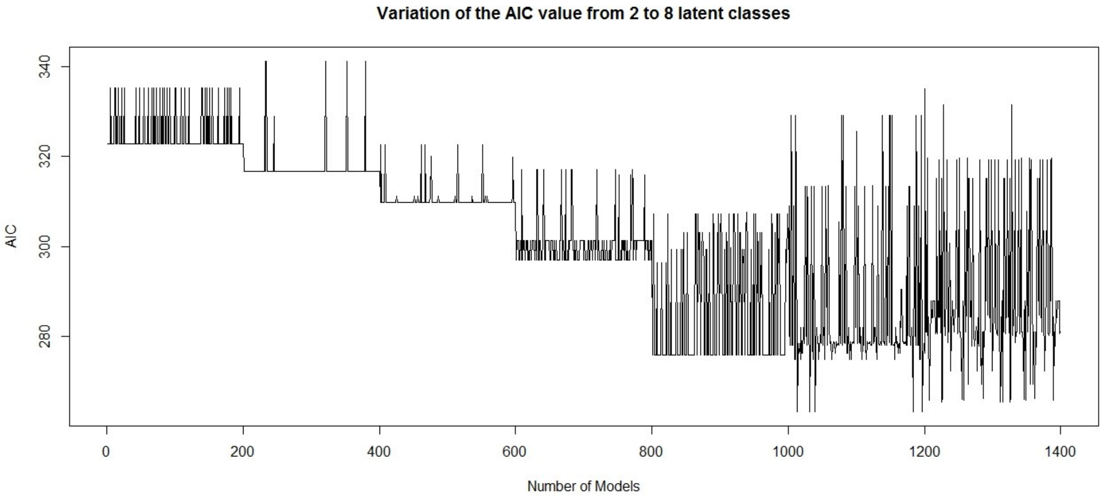
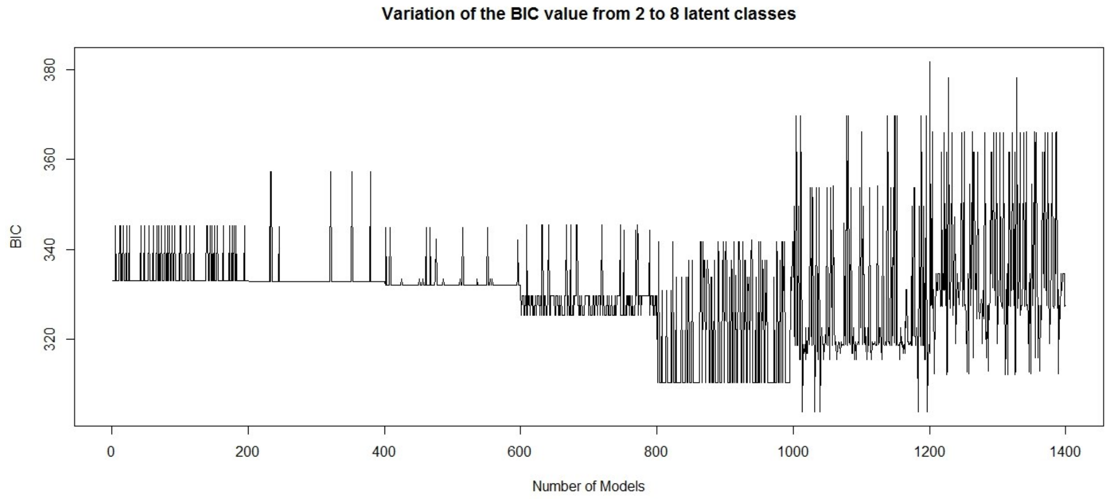


Fig. 13 BIC plot



Larger number of states can be hard to interpret. However, since the items were constructed based on the MHC, it was predicted that its difficulties would be explained by seven states. The response model coefficient for each latent class matched exactly the mean difficulty of each cluster of item's difficulty: -6.61 for the pre-operational items, -4.61 for the primary items, -1.62 for the concrete items, 0.31 for the abstract items, 1.75 for the formal items, 3.58 for the systematic items and 7.19 for the metassystematic items. So, it can be concluded that the seven-class model represents the expected seven stages of items difficulties.

DISCUSSION

The current study aimed to check for developmental stages of reasoning, studying the structural validity of the Inductive Reasoning Developmental Test (IRDT) 3rd version. Since the items were designed based on the MHC, it was expected that each group of eight items constructed with the (hypothesized) same hierarchical complexity would form clusters in terms of difficulty (prediction 1) and the mean difficulty of each cluster would present a statistically significant difference from the next adjacent cluster (prediction 2). Seven first-order factors was expected to explain each group of eight items with the same hierarchical complexity (prediction 3), a general second-order factor would explain the seven first-order latent variables (prediction 4), and one first-order general factor were not expected to explain the 56 observables variables (prediction 5). Finally, it was predicted that, since the instrument was constructed to identify seven different developmental stages, the item's difficulty distribution would be explained by seven latent classes (prediction 6).

The results showed that neither prediction can be refuted. The 56 IRDT's items fitted the dichotomous Rasch model (Infit mean = .96; SD = .17) with a high reliability estimate

(1.00), and their difficulty distribution formed a clear seven cluster structure with gaps between them (see figure 10). Differences between mean difficulties of item's clusters were statistically significant in the 95% confidence interval level, as verified through one-sample t-test. The principal contrast analysis' result suggested the unidimensionality of the items, since the percentage of raw variance explained by the measures (modeled) is moderately high (70.3%), and the residual's unexplained variance was 5.6% for the first contrast. Similar results were found by Golino, Gomes, Commons and Miller (in press), using the IRDT 1st and 2nd versions. These findings corroborate predictions 1 and 2. The previous versions of the IRDT presented a relevant issue, since the abstract, formal and systematic items were dependent on a reference table. The version used in the current study modified these items in order to solve the issue, and also introduced eight new items constructed to identify the metassystematic stage. This stage was introduced to extend the up end of the latent variable, since many participants have had maximum score on the previous version.

The use of the Rasch models in developmental stage data has been reported in previous studies (Commons et al., 2008; Bond & Fox, 2001; Dawson, 2000, 2002; Dawson, Xie & Wilson, 2003; Dawson-Tunik, 2004; Dawson-Tunik, Commons, Wilson & Fischer, 2005; Dawson-Tunik et al., 2010; Demetriou & Kyriakides, 2006; Müller, Sokol, & Overton, 1999). Among its benefits, it can be pointed that the Rasch models doesn't need a representative sample for unbiased item estimates, a norm group for comparison between individuals, giving meaning to the scores, and a normally distributed score for achieving interval scales properties (Embreston & Reise, 2000). As pointed by Andrich (2004) the Rasch models "… arises from a mathematical formalization of invariance which also turns out to be an operational criterion for fundamental measurement" (p.15). So, instead of data modeling, the Rasch's paradigm focuses on the verification of data fit to a fundamental measurement criterion, compatible with those found in the physical sciences (Andrich, 2004. p.15). So, the use of the Rasch family of statistical models help the construction of objective and additive scales, with equal-interval properties (Bond & Fox, 2001; Embreston & Reise, 2000), producing linear measures, giving estimates of precision, allowing the detection of misfit, enabling the parameters' separation of the object being measured and of the measurement instrument (Panayides, Robinson & Tymms, 2010) as well as the verification of hierarchical sequences of both item and person (Dawson, Xie & Wilson, 2003).

The studies using the Rasch models reported in table 1 show a high reliability of items and an adequate fit to the models employed. Evidences of developmental stages are verified through the distribution of items difficulties along the latent variable (Dawson, 2000; Dawson, Xie, & Wilson, 2003; Bond & Fox, 2001; Müller, Sokol, & Overton, 1999), through the categories' characteristics curves (Dawson-Tunik, 2004; Dawson-Tunik, Commons, Wilson & Fischer, 2005), using univariate statistics such as t-tests (Bond & Fox, 2001; Commons et al., 2008; Dawson, 2002; Golino, Gomes, Commons, & Miller, in press) and applying latent class analysis (Bond & Fox, 2001; Dawson-Tunik et. al., 2010; Demetriou & Kyriakides, 2006).

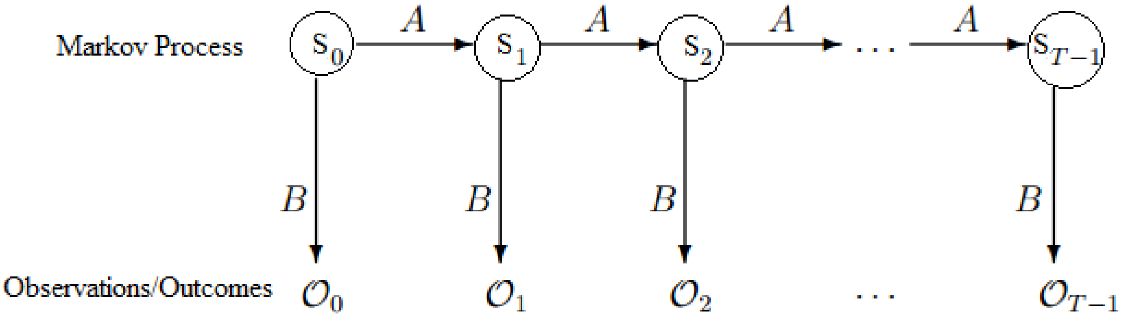
The use of different techniques and methodologies helps increasing stage evidences' strength. Demetriou and Kyriakides (2006), for example, employed the CFA to verify the structure of an intelligence battery, and the Rasch model, a cluster analysis and the Saltus model to uncover successive developmental stage-like levels of difficulty. In the present study the CFA was also used to identify the structure of the IRDT 3rd version, but instead of seeking validity evidences for different domains, as Demetriou and Kyriakides (2006), we were investigating the difficulty structure of only one domain, i.e. inductive reasoning. The result of the current study pointed to the rejection of a first-order general factor ($\chi2$ (61) = 8832.594, p = .000, CFI = .885, RMSEA = .105), and to a non-rejection of a seven first-order latent variables, each one representing a developmental stage, plus a second-order general factor (χ^2 (61) = 8832.594, p = .000, CFI = .96, RMSEA = .059) corroborating both the unidimensionality and the difficulty clusters found in the Rasch analysis, although each method investigates different aspects of data-structure (Ewing, Salzberger, & Sinkovics, 2005). These findings corroborate predictions 3, 4 and 5.

In spite of being a robust method to verify discontinuity in developmental data, having the merit of testing if "the difficulty of a group of items is significantly different for groups of persons who have different ability estimates" (Dawson-Tunik, et al., 2010, p. 06), in a mixture extension of the Rasch Model, the Saltus Model (citar) does not allow for detection of the number of latent classes explaining the distribution of items difficulties. Since the focus of the current paper is the identification of stages of item's difficulty, a more general latent class analysis was preferred. The result of the present paper indicates seven well-separated latent classes explaining the distribution of IRDT's item difficulties. Each response model coefficients, for every latent class, matched the mean Rasch difficulty estimates of each group (cluster) of items. It means that each latent class of the resulting model is a particular predicted stage. These findings corroborate prediction 6.

In sum, the current study presented evidences of the IRDT' structural validity, by showing adjust of all items to the dichotomous Rasch model, with high reliability, and evidences of unidimensionality. As predicted by theory the items presented seven clusters, visually verified in the Wright map, with significant differences between their means (95% confidence interval). Seven first-order factors explain the observable variables, and are explained by a second-order general factor. Applying the LCM on items' difficulties resulted in a model with seven well-separated classes. These findings points to developmental stage's evidence, using different methods.

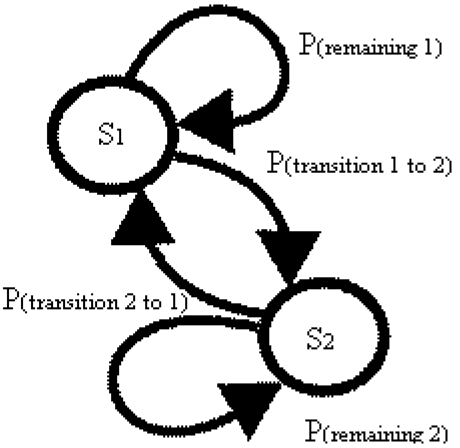
Future researches should benefit from increasing the number of adults and elderly people. Also, it would be valuable to investigate developmental stages of people employing the Saltus model, and to verify how stage transition works, which is one of the main issues of the developmental stages field. In order to study stage transition, there is an extension of the latent class model, called hidden Markov Model (HMM), which can help future researches to better understand the development of human reasoning through different stages. The HMM is based on two assumptions: 1) the current state depends only on the previous state (first-order Markov Process), and 2) observable outcomes are dependent only on the current state, at time *t*. The subjacent logic of the HMM is very close to the idea of developmental stages, in which the sequence is ordinal and not arbitrary, i.e. stage $\boldsymbol{S_n}$ is followed by a higher stage $\boldsymbol{S_{n+1}}$, and the performance of the person is related to the level of complexity of tasks (Commons, 2008), depending on the current ability level. In other words a given developmental stage depends on the previous developmental stage, and the outcome of a person in a task is attached to his current stage of performance. Figure 14 below illustrates a HMM:

Fig. 14 Hidden Markov Model



Where S_n represents states sequence from 0 to T-1, A represents state transition probabilities, B observation/outcome matrix probabilities and O_n observation/outcome sequence from 0 to T-1. Transition probabilities between states are assumed to follow first-order Markov process, i.e. state at current time depend on the previous state (Visser & Speekenbrink, 2010). Every state has a probability of remaining unchangeable, and a probability of transiting for any other state (see figure 15). This particular characteristic of the HMM is pretty relevant for the developmental stages literature, since it is a robust way to verify the developmental sequence of stages and the size of the gaps, or spacing, between stages. Although being a good method to empirically verify the sequence of stages (especially on items), the Rasch model can provide little information regarding stage spacing. Applying a probabilistic model that can give us estimates of transition between stages can be a plausible way to tackle this question, but would demand a repeated measurement design.

Fig. 15 Hidden Markov Model – transition probabilities.



REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (p. 267- 281). Budapest, Hungary: Akademai Kiado.

Andrich, D. (2004). Controversy and the Rasch model: a paradigm of incompatable paradigms. *Medical Care*, *42*(1).

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytical studies*. New York: Cambridge University Press.

Commons, M. L. (2008). Introduction to the model of hierarchical complexity and its relationship to postformal action . *World Futures*, *64*, 305-320.

Commons, M. L., Pekker, A. (2008). Presenting the formal theory of hierarchical complexity. *World Futures*, *64*, 375-382.

Commons, M., Goodheart, E., Pekker, A., Dawson, T., Draney, K., & Adams, K. (2008). Using Rasch scaled stage scores to validate orders of hierarchical complexity of balance beam task sequences. *Journal of Applied Measurement*, 9(2), 182-199.

Commons, M., Goodheart, E., Pekker, A., Dawson, T., Draney, K., & Adams, K. (2008). Using Rasch scaled stage scores to validate orders of hierarchical complexity of balance beam task sequences. *Journal of Applied Measurement*, 9(2), 182-199.

Commons, M.L., Gane-McCalla, R., Barker, C.D. & Li, E.Y (in press). The Model of Hierarchical Complexity as a Measurement System. *Journal of Adult Development*. Available at http://adultdevelopment.org/jad_special_issue.php.

Dawson, T. L. (2000). Moral reasoning and evaluative reasoning about the good life. *Journal of Applied Measurement, 1*, 372-397.

Dawson, T. L. (2002). New tools, new insights: Kohlberg's moral reasoning stages revisited. *International Journal of Behavioral Development, 26,* 154-166.

Dawson, T. L., Xie, Y., & Wilson, M. (2003). Domain-general and domain-specific developmental assessments: Do they measure the same thing? *Cognitive Development*, *18,* 61-78.

Dawson, T., Goodheart, E., Draney, K., Wilson, M., & Commons, M. (2010). Concrete, abstract, formal, and systematic operations as observed in a 'Piagetian' balance-beam task series. *Journal of Applied Measurement*, 11(1), 11-23.

Dawson, T., Goodheart, E., Draney, K., Wilson, M., & Commons, M. (2010). Concrete, abstract, formal, and systematic operations as observed in a 'Piagetian' balance-beam task series. Journal of Applied Measurement, 11(1), 11-23.

Dawson-Tunik, T. L. (2004). "A good education is…" The development of evaluative thought across the life-span. *Genetic, Social, and General Psychology Monographs, 130,* 4 112.

Dawson-Tunik, T. L., Commons, M., Wilson, M., & Fischer, K. (2005). The shape of development. *The European Journal of Developmental Psychology, 2,* 163-196.

Demetriou, A., & Kyriakides, L. (2006). The functional and developmental organization of cognitive developmental sequences. *British Journal of Educational Psychology*, 76(2), 209 242.

Dziak, J.J., Coffman, D.L., Lanza, S.T., & Li, R. (2012). Correctional populations of the United States, 2002. (Report No. #12-119). Retrieved from The Pennsylvania State University, College of Health and Human Development, The Methodology Center website: http://methodology.psu.edu/media/techreports/12-119.pdf.

Embretson, S.E. and Reise, S. P. (2000). *Item response theory for psychologists.* London: Erlbaum.

Ewing, M. T., Salzberger, T., & Sinkovics, R.R. (2005). An Alternate Approach to Assessing Cross-Cultural Measurement Equivalence in Advertising Research. *Journal of Advertising, 34*(1), 17-36.

Fischer, K. W., & Bidell, T. R. (1998). Dynamic development of psychological structures in action and thought. In W. Damon, R. M. Lerner, W. Damon, R. M. Lerner (Eds.), *Handbook of child psychology: Volume 1: Theorectical models of human development (5th ed.)* (pp. 467-561). Hoboken, NJ US: John Wiley & Sons Inc.

Fischer, K. W., & Bidell, T. R. (2006). Dynamic development of action, thought, and emotion. In W. Damon & R. M. Lerner (Eds.), *Theoretical models of human development. Handbook of child psychology* (6th ed., Vol. 1, pp. 313-399). New York: Wiley.

Fischer, K. W., & Yan, Z. (2002a). The development of dynamic skill theory. In R. Lickliter & D. Lewkowicz (Eds.), *Conceptions of development: Lessons from the laboratory*. Hove, U.K.: Psychology Press.

Fischer, K. W., & Yan, Z. (2002b). Darwin's construction of the theory of evolution: Microdevelopment of explanations of variation and change of species. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition Processes in Development and Learning*. Cambridge, U.K.: Cambridge University Press.

Fischer, K.W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87*, 477-531.

Fischer, K.W., & Dawson, T. L. (2002). A new kind of developmental science: Using models to integrate theory and research. *Monographs of the Society for Research in Child Development*, *67* (1), 156-167.

Fischer, K.W., & Rose, S.P. (1994). Dynamic development of coordination of components in brain and behavior: A framework for theory and research. In G. Dawson & K.W. Fischer (Eds.). *Human behavior and the developing brain* (pp. 3-66). New York: Guilford Press.

Fischer, K.W., & Rose, S.P. (1999). Rulers, models, and nonlinear dynamics: measurement and method in developmental research. In G. Savelsbergh, H. van der Maas, and P. van Geert (Eds.), *Nonlinear developmental processes* (pp. 197-212).

Fischer, K.W., & Silvern, L. (1985). Stages and individual differences in cognitive development. *Annual Review of Psychology, 36*, 613-648.

Fischer, K.W., Kenny, S.L., & Pipp, S.L. (1990). How cognitive processes and environmental conditions organize discontinuities in the development of abstractions. In C.N. Alexander, E.J. Langer, & R.M. Oetzel (Eds.), *Higher stages of development*. New York: Oxford University Press. Pp. 162-187.

Glas, C.A. (2007). *Multivariate and Mixture Distribution Rasch Models*. New York: Springer-Verlag.

Golino, H.F., Gomes, C.M.A., Commons, M.L.C., and Miller, P. M. (in press). The Construction and Validation of a Developmental Test for Stage Identification: Two Exploratory Studies. *Journal of Adult Development*. Available at http://adultdevelopment.org/jad_special_issue.php.

Gomes, C. M. A. & Borges, O. N. (2009). Qualidades Psicométricas do Conjunto de Testes de Inteligência Fluida. *Avaliação Psicológica*, *8*, 17-32.

Gomes, C. M. A. (2010). Estrutura fatorial da Bateria de Fatores Cognitivos de Alta-ordem (BAFACALO). *Avaliação Psicológica, 9*, 449-459.

Gomes, C.M.A. & Golino, H.F. (2009). Estudo exploratório sobre o Teste de Desenvolvimento do Raciocinio Indutivo (TDRI). In D. Colinvaux. *Anais do VII Congresso Brasileiro de Psicologia do Desenvolvimento: Desenvolvimento e Direitos Humananos.* (pp. 77-79). Rio de Janeiro: UERJ. Available at http://www.abpd.psc.br/files/congressosAnteriores/AnaisVIICBPD.pdf.

Haughton, D., Legrand, P., & Woolford, S. (2009). Review of Three Latent Class Cluster Analysis Packages: Latent Gold, poLCA, and MCLUST. *The American Statistician, 63*(1), 81-91.

Hibbard, J., Collins, P., Mahoney, E., & Baker, L. (2010). The development and testing of a measure assessing clinician beliefs about patient self-management. *Health Expectations: An International Journal of Public Participation in Health Care & Health Policy*, *13*(1), 65 72.

Inhelder, B., and Piaget, J. (1958). *The growth of logical thinking forom child-hood to adolescence.* New York: Basic Books.

Linacre J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16* (2), 878.

Linacre J. M. (2011). WINSTEPS. Rasch measurement computer program, Winsteps.com, Chicago.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 32*(2), 103-122.

Muthén, B. (1989). Factor structure in groups selected on observed scores. British Journal of *Mathematical and Statistical Psychology, 42*, 81-90. (#23)

Panayides, P., Robinson, C., & Tymms, P. (2010). The assessment revolution that has passed England by: Rasch measurement, *British Educational Research Journal*, *36*(4), 611 626.

Peeters, M.J. & Stone, G.E. (2009). An Instrument to Objectively Measure Pharmacist Professionalism as an Outcome: A Pilot Study. *The Canadian Journal of Hospital Pharmacy, 62*(3), 209-216.

Rose, S. P., & Fischer, K. W. (1998). Models and rulers in dynamical development. *British Journal of Developmental Psychology*, *16*(1), 123-131.

Schwartz, M. S., & Fischer, K. W. (2005). Building general knowledge and skill: Cognition and microdevelopment in science learning. In A. Demetriou & A. Raftopoulos (Eds.), Cognitive developmental change: Theories, models, and measurement. Cambridge, U.K.: Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6* , 461-464.

Stein, Z., & Hiekkinen, K. (2009). Metrics, models, and measurement in developmental psychology. *Integral Review, 5*(1), 4-24.

Tennant A. & Pallant J.F. (2006), Unidimensionality Matters! (A Tale of Two Smiths?) *Rasch Measurement Transactions*,*20*(1), 1048-51.

Van Geert, P. & Steenbeek, H. (2005). Explaining after by before: Basic aspects of a dynamic systems approach to the study of development. *Developmental Review*, 25, 408 442.

Visser, I. & Speekenbrink, M. (2010). depmixS4: An R Package for Hidden Markov Models. *Journal of Statistical Software, 36*(7), 1-21. URL http://www.jstatsoft.org/v36/i07/.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276-289.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Yan, Z., & Fischer, K. W. (2007). Pattern emergence and pattern transition in microdevelopmental variation: Evidence of complex dynamics of developmental processes. *Journal of Developmental Processes*, *2*(2), 39-62.

3. Conclusão

A presente dissertação teve como objetivo verificar a validade de estágios de desenvolvimento do raciocínio indutivo, por meio da construção e validação do Teste de Desenvolvimento do Raciocínio Indutivo (TDRI). Ela foi dividida em dois artigos. O primeiro apresentou as duas versões iniciais do TDRI, e investigou se os itens mensuram os estágios de desenvolvimento, formando grupamentos distintos entre si, em duas amostras, uma composta por 167 pessoas (50.3% homens) com idades entre 6 e 58 anos (M = 18,90, DP = 9,70), e a outra composta por 188 pessoas (57.7% mulheres) com idades entre 6 e 65 anos (M = 21,45, DP = 14,31). Os resultados apontaram um adequado ajuste ao modelo dicotômico de Rasch (infit médio = 0,94; desvio-padrão = 0,22), e evidenciaram que os grupamentos de itens seguem o padrão previsto (oito itens por grupamento, cada grupamento formando um estágio), e que grupamentos adjacentes apresentam diferenças significativas entre si. O segundo artigo investigou a validade estrutural da 3ª versão do TDRI, que foi construída para superar algumas limitações verificadas nas primeiras duas versões. Esse segundo estudo empregou três metodologias distintas para verificar a validade dos estágios de desenvolvimento: 1) Análise Fatorial Confirmatória (AFC); 2) Análise Rasch para dados dicotômicos; e 3) Análise de classes latentes. A amostra foi composta por 1.459 pessoas people (52.5% mulheres) com idade entre 5 e 86 anos (M = 15,75, DP = 12,21). O resultado apontou uma estrutura fatorial de dois níveis, sendo o primeiro nível composto por 7 fatores (um para cada estágio) e o segundo nível um fator geral [$\chi 2$ (61) = 8832.594, p = .000, CFI = .96, RMSEA = .059]. Os 56 itens do TDRI 3ª versão se ajustaram ao modelo Rasch (infit médio = 0,96; desvio-padrão = 0,17), e apresentaram uma confiabilidade alta para os itens (1.00) e moderadamente alta para as pessoas (0,82). As evidências apontaram que a solução com sete classes latentes apresenta o melhor ajuste aos dados (AIC: 263.380; BIC: 303.887; Loglik: -111.690). Os estudos que compõe essa dissertação mostram que é possível, a partir da adoção de uma série de metodologias específicas, identificar empiricamente estágios de desenvolvimento. As evidências apontam que o TDRI é um instrumento válido e confiável para avaliar estágios de desenvolvimento do raciocínio indutivo.