© 2014 AMERICAN PSYCHOLOGICAL ASSOCIATION

BEHAVIORAL DEVELOPMENT BULLETIN

VOLUME 19
NUMBER 3
SEPTEMBER 2014

The construction and validation of a developmental test for stage identification: Two exploratory studies

Hudson F. Golino[1], Cristiano Mauro Assis Gomes[1], Michael Lamport Commons[2], and Patrice Marie Miller[3]

1 Universidade Federal de Minas Gerais, Brazil
2 Harvard Medical School, USA
3 Salem State University, USA

ABSTRACT

The present work presents two exploratory studies about the construction and validation of the Inductive Reasoning Developmental Test (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. The one sample *t*-tests showed significant spacing of Rasch scores between items of adjacent orders of hierarchical complexity, 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 and 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. The one sample *t*-tests showed significant spacing of Rasch scores between items of adjacent orders of hierarchical complexity, 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

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 this 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 (Marshall, 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 works (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 the developmental stages of inductive reasoning through reliable, valid and accurate measures, falling in the category of so-called "quality control standards".

Author note: 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. Special thanks to 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 and commentaries on the original manuscript.

Correspondence regarding this article should be addressed to Hudson F. Golino, Laboratory for Cognitive Architecture Mapping (LaiCo), Universidade Federal de Minas Gerais, Brasil. E-mail: hfgolino@gmail.com. Mobile: + 55 31 88607490

Criticisms of stages, or killing Piagetian stage theory:

Starting 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 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 in 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 suggests 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 already in his early work entitled *The Developmental Psychology of Jean Piaget* (1963), pointed to ambiguities in the concept of stage, argued about the challenges of the clinical method, on the impossibility of stating that a child "has" a particular concept and raised 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, Flavell 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 (Marshall, 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 modern theoretical and methodological approaches that have been providing new evidences for discontinuity. Included in these new approaches are two important and related models of development: Fischer's Dynamic Skill Theory (DST; Fischer, 1980; 2008) and Commons' Model of Hierarchical Complexity (MHC; Commons, 2008; Commons & Richards, 1984; Commons & Pekker, 2008). 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, 2002b) 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) and 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 provided by Fischer and colleagues (Fischer, Kenny, & Pipp, 1990; Fischer & Silvern, 1985; Fischer & Yan, 2002a, 2002b; Schwartz & Fischer, 2005; Yan & Fischer, 2007). Instead of conceptualizing the discontinuous facet of human development as a unidirectional ladder the DST sees it as a *constructive web* that encompasses someone's activity and the supportive context in which this activity 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 *decaláge* 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, it requires the two actions of addition and multiplication to be performed in a certain order, thus, coordinated.

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)

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). 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.

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 it 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, 1987b) 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 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.

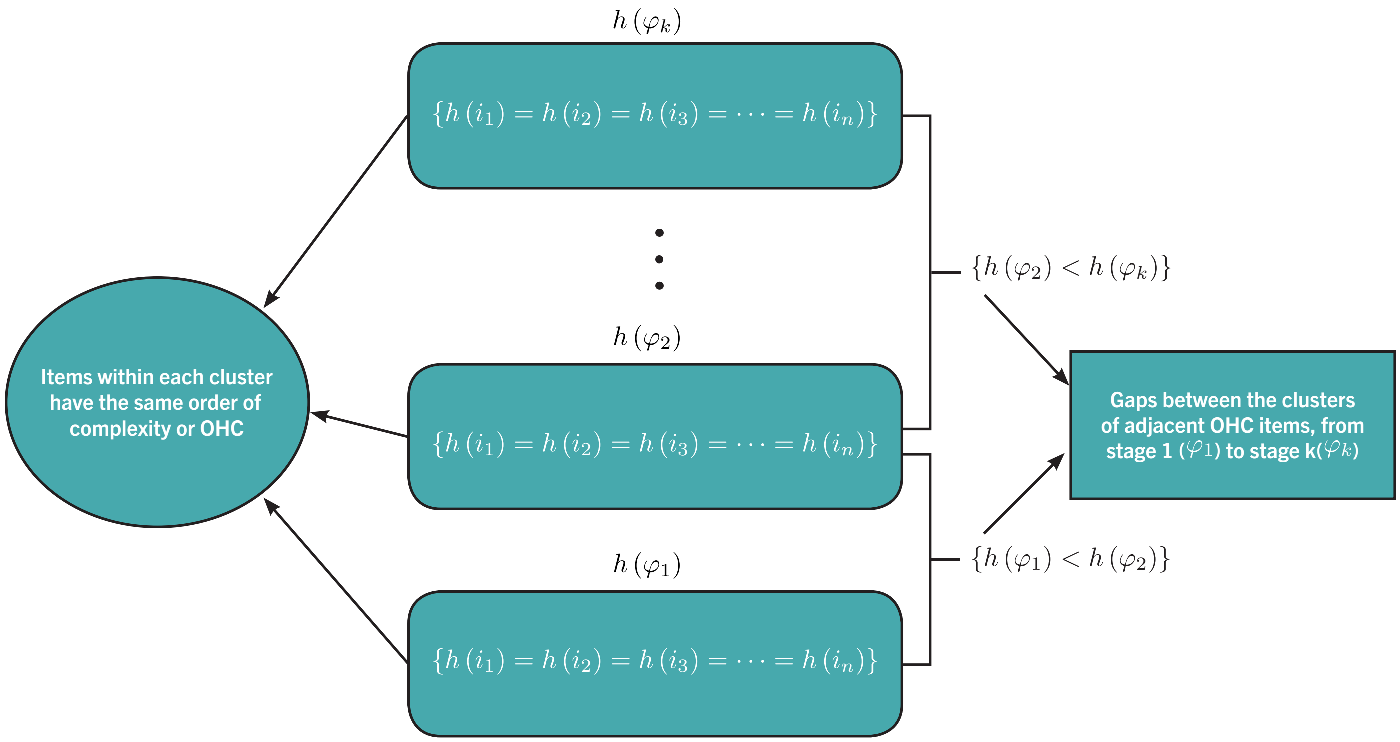

Figure 1. Expected item structure of instruments constructed focusing on the vertical complexity within a specific domain (unidimensional)

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

Constructing calibrated tests for developmental stage identification requires a specific design as defined by Commons and colleagues (Commons & Pekker, 2008; Commons, Gane-McCalla, Barker & Li, this issue). This design involves: 1) grouping items with same hierarchical complexity [$h(i_1) = h(i_2) = h(i_3) = \ldots h(i_n)$] within stages; and 2) using items with increasing hierarchical complexity [h(Stage 1) < h(Stage 2) < h(Stage 3) < … h(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; see Commons & Pekker, 2008; Commons, Gane-McCalla, Barker & Li, this issue). Furthermore, the figure shows the expected gaps between the clusters of adjacent OCH items (see Figure 1).

Beyond the strategies of grouping items with same OHC and using items with increasing OHC, in order to identify developmental stages, 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 et al., 2008; Dawson, Goodheart, Draney, Wilson, & Commons, 2010) evaluated 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:

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

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 et al., 2008; Dawson-Tunik et al., 2010) 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 discontinuity. 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.

1	A	A	A	A	E

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

The IRDT (Gomes & Golino, 2009) is a pencil-and-paper instrument designed 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).

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

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

13	XZ	DF	MO	RU	HJ

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

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).

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).

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).

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).

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).

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).

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

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).

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).

20	KLNO	NPQR	QSTU	DFGH	HJKL

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

17	IJKM	NOPR	EFHI	QRSU	JKLN

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

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. It is important to note that the revision of the MHC stages was not incorporated in this study. The OHC numbers presented in this study are the older version of MHC stages. In the newer version, the numbers increase by one.

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).

» METHOD

Participants

In Study 1, the IRDT was administered to a convenience sample composed by 167 Brazilians (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, 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 to administer the IRDT to three different people from 6 to 60 years of age. Participation was voluntary, with people agreeing to be part of the study after its purpose 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

22	RTVW	ACEF	CEGH	FHJK	BDEG

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

Table 2. One-sample tests of mean item difficulties for different OHC's

	Test value = 0								
						95% CI		Effect size	
Stages	*t*	*DF*	Sig. (2-tailed)	*M*	*SD*	Lower	Upper	(*d'*)	(*r*)
Pre-op/SR and Primary/RM	13.58	7	0.00	3.82	0.80	3.15	4.48	10.26	.98
Primary/RM and Concrete/RS	3.29	7	0.01	2.18	1.87	0.61	3.74	2.49	.77
Concrete/RS and Abstract/SA	7.99	7	0.00	1.69	0.60	1.19	2.18	6.03	.94
Abstract/AS and Formal/AM	36.01	7	0.00	2.89	0.23	2.70	3.08	27.28	.99
Formal/AM and Systematic/AS.	9.49	7	0.00	2.28	0.68	1.71	2.85	7.17	.96

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:

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

where $\overline{X}_{k+1}$ is the mean of the next higher order of complexity (or Stage *k*+1), and δ_{i_k} 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 (ES) is calculated using the Cohen's *d* and effect size correlation *r* (Rosnow & Rosenthal, 1996).

» 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 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 14) 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).

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

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

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

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

25	Aπ	Aδ	Aη	Aμ	Aλ

Figure 10. Example: Item 25, Abstract/SA

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, effect size *d'* and *r* were large (see Table 2).

» 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.1).

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

41	AδBηFπAμ	CηEπBλDδ	AδFπAμDδ	BπFλCδFη	BδEλAπFμ

Figure 12. Example: Item 41 ,Systematic/AS

33	AδFπ	CηEπ	BδEλ	CδFη	AλBπ

Figure 11. Example: Item 33, Formal/AM

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.

Table 3. One-sample tests of mean item difficulties for different OHC's

	Test value = 0								
						95% CI		Effect size	
Stages	*t*	*DF*	Sig. (2-tailed)	*M*	*SD*	Lower	Upper	(*d'*)	(*r*)
Pre-op/SR and Primary/RM	10.36	7	0.00	3.61	0.99	2.79	4.43	7.83	.96
Primary/RM and Concrete/RS	22.94	7	0.00	3.42	0.42	3.06	3.77	17.34	.99
Concrete/RS and Abstract/SA	23.03	7	0.00	3.33	0.41	2.99	3.67	17.40	.99
Abstract/AS and Formal/AM	10.96	7	0.00	1.14	0.29	0.89	1.38	8.28	.97
Formal/AM and Systematic/AS.	4.78	7	0.00	0.88	0.52	0.44	1.31	3.61	.87

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.

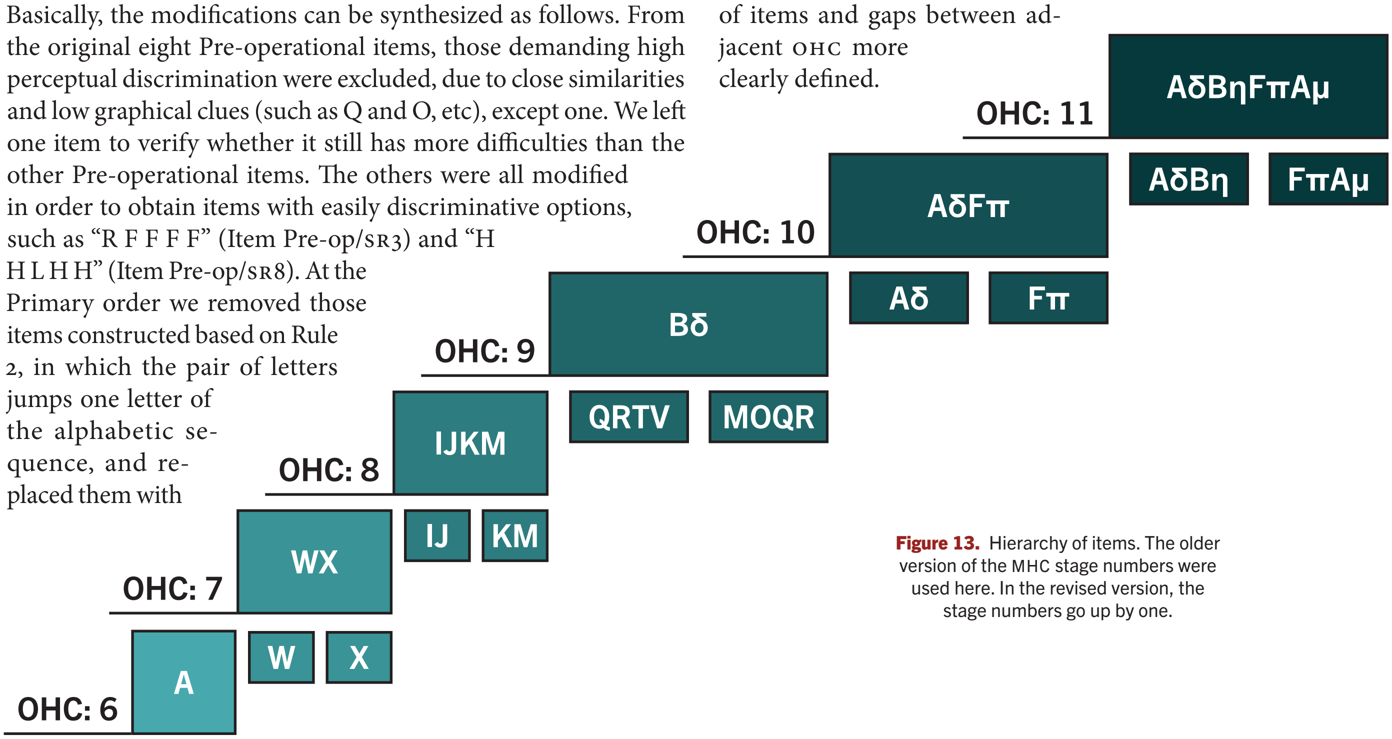

Figure 13. Hierarchy of items. The older version of the MHC stage numbers were used here. In the revised version, the stage numbers go up by one.

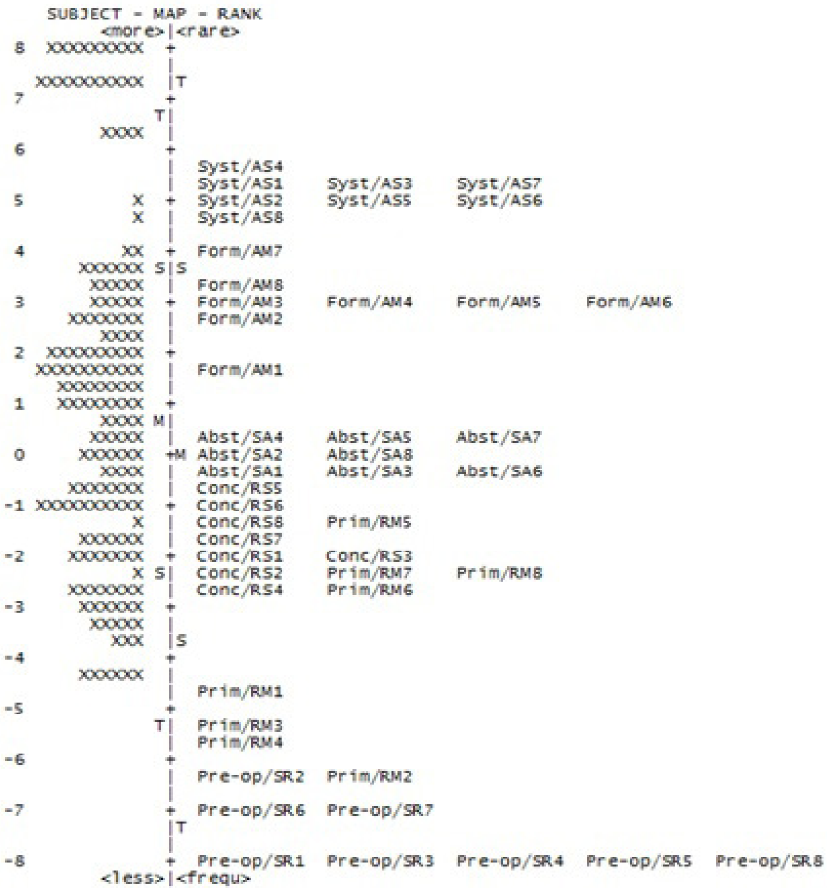

Figure 14. Variable Map showing the IRDT's items

» 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 2) 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).

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 size d' and r were large (see Table 3).

» 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:

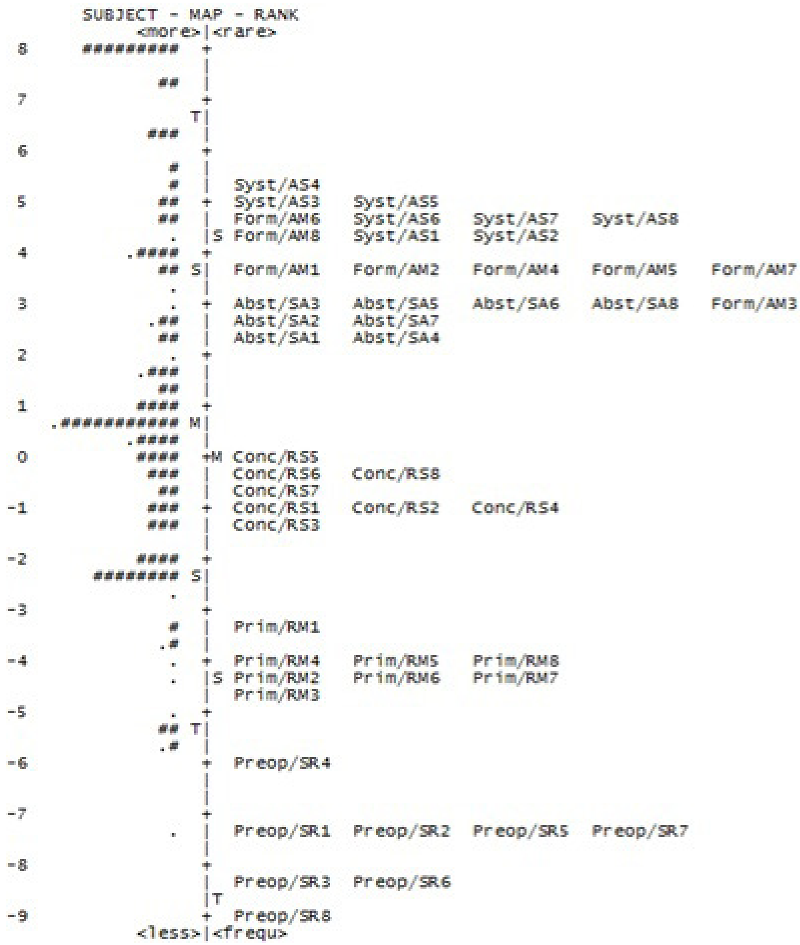

Figure 15. Variable Map showing the IRDT 2nd version's items

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

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 16 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).

» CONCLUSION

In line with previous researches (Bond & Fox, 2001; Commons et al., 2008; Dawson, 2000, 2002; Dawson, Xie, & Wilson, 2003; Dawson-Tunik, 2004; Dawson-Tunik, Commons, Wilson & Fischer, 2005), the current study adds supportive evidence for developmental stages using modern quantitative methods and a specific test design provided by the model of Hierarchical Complexity and by

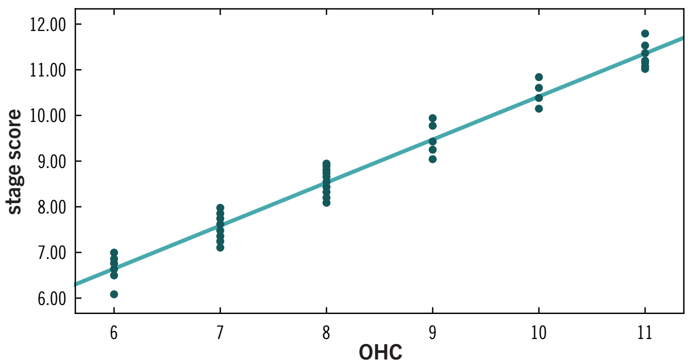

Figure 16. Regression of Stage Scores on Order of Hierarchical Complexity. The older version of the MHC stage numbers was used here. In the revised version, the stage numbers go up by one

the Dynamic Skill Theory. This study also 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 importance, 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 and apply latent class analysis (Bond & Fox, 2001; Dawson-Tunik et. al., 2010; Demetriou & Kyriakides, 2006) of item's difficulties, since the use of different quantitative techniques can benefit stage evidences' strength. ■

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), 7-16.

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*(4), 433-453.

Baillargeon, R. (1987). Object permanence in 3 1/2- and 4 1/2-month-old infants. *Developmental Psychology, 23* (5), 655-664.

Bart, W. (1971). The effect of interest on horizontal decaláge at the stage of formal operations. *Journal of Psychology: Interdisciplinary and Applied*, *78*(2), 141-150. doi:10.1080/00223980.1971.9916897

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 decaláge in the development of transitivity. *Developmental Psychology*, *24*(4), 542-551. doi:10.1037/0012-1649.24.4.542

Chazan, S. (1972, May). Horizontal decaláge in the concept of object permanence as a correlate of dimensions of maternal care. *Dissertation Abstracts International*, 32. 6613.

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*(5-7), 305-320. doi:10.1080/02604020802301105

Commons, M.L., Gane-McCalla, R., Barker, C.D., & Li, E.Y (2014). The Model of Hierarchical Complexity as a Measurement System. *Behavioral Development Bulletin, 19*(3).

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. doi:10.1016/0160-2527(95)00017-C

Commons, M. L., Pekker, A. (2008). Presenting the formal theory of hierarchical complexity. *World Futures*, *64*(5-7), 375-382. doi:10.1080/02604020802301204

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*(4), 323-340. doi:10.1016/0273-2297(90)90017-X

Commons, M. L., Rodriguez, J. A. (1993). The development of hierarchically complex equivalence classes. *Psychological Record, 43*(4), 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*(6), 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*(4), 1058-1069. doi:10.1111/1467-8624.ep8587686

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. doi:10.1006/drev.1998.0467

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(*4), 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*(2)*,* 154-166. doi:10.1080/01650250042000645

Dawson, T. L. (2003). A stage is a stage is a stage: A direct comparison of two scoring systems. *Journal of Genetic Psychology, 164*(3), 335-364.

Dawson, T. L. (2004). Assessing intellectual development: Three approaches, one sequence. *Journal of Adult Development, 11*(2)*,* 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*(3/4), 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*(1)*,* 61-78. doi:10.1016/S0885-2014(02)00162-4

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*(1)*,* 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. doi:10.1348/000709905X43256

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*(1), vii-154. doi:10.1111/1540-5834.671173

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*(3), 480-497.

Demetriou, A., Mouyi, A., & Spanoudis, G. (2008). Modelling the structure and development of g. *Intelligence*, *36*(5), 437-454. doi:10.1016/j.intell.2007.10.002

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*(3), 175-231. doi:10.1016/j.newideapsych.2004.11.005

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*(6), 477-531. doi:10.1037/0033-295X.87.6.477

Fischer, K.W. (1987). Relations between brain and cognitive development. *Child Development, 58*(3), 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. doi:10.1111/1540-5834.671184

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: Royal Netherlands Academy of Arts and Sciences.

Fischer, K.W., & Silvern, L. (1985). Stages and individual differences in cognitive development. *Annual Review of Psychology, 36*, 613-648. doi:10.1146/annurev.ps.36.020185.003145

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,* (pp. 162-187). New York: Oxford University Press.

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. (1996, June). *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. (1997, March). *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. doi:10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care*, 38(Suppl9), II60-II65. doi:10.1097/00005650-200009002-00009

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. doi:10.1111/j.2044-835X.1998.tb00751.x

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 decaláge between object permanence and person permanence. *Developmental Psychology, 14*(1). 1-10. doi:10.1037/0012-1649.14.1.1

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. (2014). Developmental stage of performance in reasoning about bullying in school age youth. *Behavioral Development Bulletin*, *19*(2), 81-99.

Kallio, E., & Helkaman, K (1991). Formal operations and postformal reasoning: A replication. *Scandinavian Journal of Psychology*, 32(1), 18-21. doi:10.1111/j.1467-9450.1991.tb00848.x

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). Switzerland: Thur AG Offsettdruck

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 Decalóge: A problem and its solution. *Cognitive Development, 4*(1)*,* 89-119. doi:10.1016/0885-2014(89)90006-3

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*), 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*. doi:10.1023/A:1016060328253*

Luce, R.D. & Tukey, J.W. (1964). Simultaneous conjoint measurement: a new scale type of fundamental measurement. *Journal of Mathematical Psychology, 1*(1), 1—27. doi:10.1016/0022-2496(64)90015-X

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. doi:10.1037/0012-1649.13.6.666

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 (2008, June). *Finding the Relationships Among Moral Development Measures Using the Model of Hierarchical Complexity and Rasch Analysis*. Paper presented at Jean Piaget Society, Quebec City, Quebec, Canada.

Miller, J. G., Harrigan, W. J., Commons, M. L. & Commons-Miller, N. H. K. (2008, November). *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. doi:10.2466/pr0.1969.25.3.941

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 decalâge. *Dissertation Abstracts International*, 31, 3347.

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. doi:10.1080/01411920903018182

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. (2008, July). *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. doi:10.1007/s11336-002-1047-z

Roberge, J. (1976). Developmental analyses of two formal operational structures: Combinatorial thinking and conditional reasoning. *Developmental Psychology*, *12*(6), 563 564. doi:10.1037/0012-1649.12.6.563

Rose, S. P., & Fischer, K. W. (1998). Models and rulers in dynamical development. *British Journal of Developmental Psychology*, *16*(1), 123-131. doi:10.1111/j.2044-835X.1998.tb00752.x

Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Pyschological Methods, 1*(4), 331-340. doi:10.1037/1082-989X.1.4.331

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. doi:10.1080/15366367.2011.558792

Scardamalia, M. (1977). Information processing capacity and the problem of horizontal Decalâge : 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. doi:10.1037/030229

Siegler, R.S, (1981). Developmental sequences within and between concepts. *Monograph of The Society for Research in Child Development*, *46*(2), 1-84. doi:10.2307/1165995

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*(3)*,* 395–417. doi:10.1037/0033-295X.99.3.395

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. doi:10.1016/j.dr.2005.10.003

Webb, R. (1974). Concrete and formal operations in very bright 6- to 11-year-olds. *Human Development*, *17*(4), 292-300. doi:10.1159/000271352

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*, Chicago: Mesa Press

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.

APPENDIX A

Description of the IRDT demands by OHC

Note. The older version of the MHC stage numbers was used here. In the revised version, the stage numbers go up by one.

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.

APPENDIX B

Inductive Reasoning Developmental Test 2nd Version

Pre-operational Items					
1	A	A	A	A	E
2	B	B	B	C	B
3	R	F	F	F	F
4	U	U	V	U	U
5	Q	Q	C	Q	Q
6	V	V	V	S	V
7	D	G	D	D	D
8	H	H	L	H	H
Primary Items					
9	WX	KL	ST	PR	YZ
10	IJ	RT	CD	UV	MN
11	TU	HI	QR	JL	BC
12	PQ	NO	GI	CD	RS
13	XY	AB	TU	DF	OP
14	ST	IK	YZ	VW	EF
15	JK	DE	UV	HI	NP
16	GH	XZ	LM	RS	KL
Concrete Items					
17	NOPR	IJKM	UVXY	MNOQ	QRSU
18	PQRT	LMNP	GHIK	VWXZ	KLNO
19	HIJL	TUWX	RSTV	OPQS	FGHJ
20	JKLN	BCDF	PQST	CDEG	STUW
21	OQST	DFHI	MOQR	EGHJ	TVXY
22	RTVW	ACEF	BDEG	CEGH	FHJK
23	IKMN	LNPQ	RTVW	JLMO	SUWX
24	GIKL	FHIK	PRTU	QSUV	CEGH

Reference Table				
Ж	Ю	ф	э	ђ
FGIKOQST	OPRTDFHI	IJLNPRSU	EFHJTVXY	RSUWNPRS
µ	π	σ	Љ	И
QRTVMOQR	STVXIKMN	KLNSUWX	CDFHNORS	GHJLPRTU
Ω	∑	∆	¥	Ħ
LMOQEGIJ	BCEGJLNO	MNPRGIKL	JKMOUWYZ	KLNPDEHI
Θ	Ξ	Π	Ψ	A
UVXKLNP	QSTVACEF	OQRTBDFG	FHIKRTVW	HJKMGIKL
œ	Ŧ	ǿ	β	δ
OPQTCEGH	JLMOPRTU	UWXZQSUV	CEFHNOPS	HJKMDFHI
Ђ	љ	ε	ζ	λ
KMNPGIKL	EGHJGHIL	QSTVMOQR	TVYKMOP	DFGISUWX
Щ	‡	þ	∫	Ґ
CDGHUVWZ	KLOPEFGJ	CDGHTUVY	LMPQDEGI	QRUVMNOR
Б	Ц	Ъ	ŧ	ŋ
TUXYIJKN	OPSTFGK	HILMNOPS	ABEFBCDG	UVYZJKMO

Abstract Items					
25	Ж	Ю	Φ	Э	ђ
26	μ	π	σ	Љ	И
27	Ω	∑	Δ	¥	Ħ
28	Θ	Ξ	Π	Ψ	α
29	œ	Ŧ	ǿ	β	δ
30	Ѣ	љ	ε	Z	λ
31	Щ	‡	Þ	∫	Ґ
32	Б	Џ	Ъ	Ŧ	ŋ
Formal Items					
33	ЮѢ	Δœ	πδ	∑ε	ђμ
34	Жζ	ЮΔ	¥α	Ωǿ	эΞ
35	ЮΩ	μλ	σŦ	ИΞ	ђΨ
36	ΨҐ	ǿ ŧ	Œλ	αЏ	Ŧþ
37	ѢЪ	δБ	λ‡	Ξ ε	Ξ Щ
38	λπ	α¥	ǿ μ	Ѣσ	δ Π
39	ε œ	œЮ	Ψђ	Ξ Ω	ε И
40	Ŧэ	ΠЖ	δ∑	ζΔ	ε Ψ
Systematic Items					
41	ЮσѢэ	Δ œ И Ξ	ЮѢэ Ξ	μ λ∑ ε	π δ Жζ
42	¥ αИ Ξ	Δ ИŦ μ	σ Ψ Δ Ŧ	π αэǿ	ђ Ψ ђ Ѣ
43	σ α∑ Π	ђ ǿ Ω δ	ЮζαΔ	¥ δ Ω Ѣ	μ Ξ∑ε
44	Π Ж ǿ μ	ǿ Ω σ œ	δ∑Ξ Ω	λπ ε И	ѢσœЮ
45	α¥ Ŧ э	λΔ ѢЖ	ǿ∑œэ	ε Ω Ξ ђ	ζπ μ α
46	Ψ ∑ μ α	ǿ ЮΨ Ω	Π эζμ	Ŧ ђ ѢИ	αΔ ǿ ¥
47	Ψ Џ Ξ ŧ	Ŧ Ґδ Џ	λþα‡	αþσ α	ǿ ЩѢҐ
48	δ БαЏ	Ψ Ґ λ‡	ǿ ŧ Ξ Щ	Ŧ þѢЪ	ζЪ∑ Π