Psychological Assessment

Developing a Fluid Intelligence Scale Through a Combination of Rasch Modeling and Cognitive Psychology

Ricardo Primi

Online First Publication, April 28, 2014. http://dx.doi.org/10.1037/a0036712

CITATION

Primi, R. (2014, April 28). Developing a Fluid Intelligence Scale Through a Combination of Rasch Modeling and Cognitive Psychology. *Psychological Assessment*. Advance online publication. http://dx.doi.org/10.1037/a0036712

Psychological Assessment

© 2014 American Psychological Association
1040-3590/14/$12.00 http://dx.doi.org/10.1037/a0036712

Developing a Fluid Intelligence Scale Through a Combination of Rasch Modeling and Cognitive Psychology

Ricardo Primi
University São Francisco

Ability testing has been criticized because understanding of the construct being assessed is incomplete and because the testing has not yet been satisfactorily improved in accordance with new knowledge from cognitive psychology. This article contributes to the solution of this problem through the application of item response theory and Susan Embretson's cognitive design system for test development in the development of a fluid intelligence scale. This study is based on findings from cognitive psychology; instead of focusing on the development of a test, it focuses on the definition of a variable for the creation of a criterion-referenced measure for fluid intelligence. A geometric matrix item bank with 26 items was analyzed with data from 2,797 undergraduate students. The main result was a criterion-referenced scale that was based on information from item features that were linked to cognitive components, such as storage capacity, goal management, and abstraction; this information was used to create the descriptions of selected levels of a fluid intelligence scale. The scale proposed that the levels of fluid intelligence range from the ability to solve problems containing a limited number of bits of information with obvious relationships through the ability to solve problems that involve abstract relationships under conditions that are confounded with an information overload and distraction by mixed noise. This scale can be employed in future research to provide interpretations for the measurements of the cognitive processes mastered and the types of difficulty experienced by examinees.

Keywords: inductive reasoning, fluid intelligence, Rasch measurement, matrix reasoning

There has been significant development in cognitive psychology and the psychometrics of intelligence testing over the last few decades (Cornoldi, 2006; Deary, 2001; Whitely, 1980; Whitely & Schneider, 1981). Cognitive task analysis of items that are commonly used in intelligence tests provides a better understanding of how people represent and process information, which in turn improves a test's score for construct validity (Sternberg, 1981). Psychometric model-based methods, including item response theory (IRT), provide ways to construct scales that show links between test scores and the underlying construct that provide substantial additional interpretations (Embretson, 2006; Wilson, 2005). These methods have recently evolved into cognitive diagnostic assessment models such as that of Tatsuoka (2009).

Despite these developments, test construction procedure has not yet satisfactorily incorporated these new methods (Embretson, 1994). Thus, the expectation for the future generation of tests is that cognitive psychology should play an important role in construct representation. By providing a rich theoretical basis for the creation of purified tasks, it should lead to the development of instruments with better theoretical grounding. This article presents an illustration of these methods as they apply to the development of a fluid reasoning test. First, a review of the psychometric and cognitive neuroscience definitions of fluid intelligence is presented. A brief review of how fluid intelligence is measured follows, with emphasis on the justification for new tests based on modern methodology. Finally, an empirical study of test development is presented.

Nature of Fluid Intelligence

According to Schneider and McGrew (2012), fluid reasoning (Gf) refers to

> the deliberate but flexible control of attention to solve novel "on the spot" problems that cannot be performed by relying exclusively on previously learned habits, schemas and scripts. Fluid reasoning is a multi-dimensional construct, but its parts are unified in their purpose: solving unfamiliar problems. Fluid reasoning is most evident in abstract reasoning that depends less on prior learning. However, it is also present in day-to-day problem solving. Fluid reasoning is typically employed in concert with background knowledge and automatized responses. (p. 111)

Fluid intelligence is central to understanding the construct of intelligence. As it is the broad factor most related to the general

This article was developed as part of a larger project, called Development of a Computerized Componential Fluid Intelligence Test, which was financed by the Foundation for the Support of Research in the State of São Paulo (FAPESP, Process 2000/05913-4), by the Brazilian National Research Council (CNPq), and by the University of São Francisco. The author acknowledges the contributions of Acácia A. Angeli dos Santos, Claudette Medeiros Vendramini, Fernanda de Oliveira Soares Taxa, Maria de Fátima Lukjanenko, Franz A. Müller, Isabel Sampaio, Fatima Keiko Kuse, and Cíntia Heloína Bueno for their valuable comments on drafts of this article and assistance in collection of the data.

Correspondence concerning this article should be addressed to Ricardo Primi, Universidade São Francisco, Laboratório de Avaliação Psicológica e Educacional, Rua Alexandre Rodrigues Barbosa, 45, CEP 13251-900, Itatiba, São Paulo, Brazil. E-mail: rprimi@mac.com

factor *g* (Blair, 2006; Schneider & McGrew, 2012), it predicts various important outcomes such as achievement and rate of learning, especially with regard to complex subjects like math (McGrew & Wendling, 2010; Primi, Ferrão, & Almeida, 2010), and job performance in highly complex occupations (Gottfredson, 1997). Fluid intelligence has recently been shown to play an important role in individual differences in creativity (Silvia & Beaty, 2012). Fluid intelligence relates to the activity of the prefrontal cortex, including related damages (Duncan, Burgess, & Emslie, 1995; Duncan, Emslie, & Williams, 1996; Duncan, Schramm, Thompson, & Dumontheil, 2012; Kane, 2005); therefore, it is potentially relevant for differential diagnosis of frontal lobe lesions. Also, carefully controlled latent variable modeling studies have revealed that the latent factor of fluid intelligence is central to explaining the construct of executive functions, which is a core part of any neuropsychological battery (Salthouse, 2005, 2011; Salthouse, Pink, & Tucker-Drob, 2008).

Over the past four decades, the basic cognitive process underlying fluid intelligence tasks has been the subject of studies by cognitive psychologists (Bethell-Fox, Lohman, & Snow, 1984; Carpenter, Just, & Shell, 1990; Embretson, 1995, 1998, 2002; Marshalek, Lohman, & Snow, 1983; Mulholland, Pellegrino, & Glaser, 1980; Primi, 1995, 1998, 2002; Sternberg, 1977, 1978; Sternberg & Gardner, 1983), cognitive neuroscientists, and neuroimaging researchers (Burgess, Gray, Conway, & Braver, 2011; Gray, Chabris, & Braver, 2003; Kane, 2005; Prabhakaran, Smith, Desmond, Glover, & Gabrieli, 1997). Most of the individual differences in fluid intelligence are shared with working memory capacity (Ackerman, Beier, & Boyle 2005; Embretson, 1995, 1998; Engle, Tuholski, Laughlin, & Conway, 1999; Kane, Hambrick, & Conway, 2005; Kyllonen & Christal, 1990; Salthouse et al., 2008). According to Baddeley and Hitch (1994), working memory capacity can be deconstructed into memory buffers that are responsible for storing speech-based information and visuospatial information (phonological loops and a visuospatial sketchpad) and a central executive component that is responsible for the coordination of the basic components and attentional control. The central executive component has been broken down into a number of interrelated functions, such as (a) the coordination of simultaneous mental activities, (b) the monitoring of mental activities, (c) attentional control and selective attention, (d) the activation of information from long-term memory, and (e) rerouting or adaptive flexibility (Shimamura, 2000). What appears to be the basic cognitive mechanism that explains this shared variance among fluid intelligence, central executive, and working memory is controlled attention (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Engle et al., 1999; Heitz, Unsworth, & Engle, 2005; Kane, Bleckley, Conway, & Engle, 2001; Miyake, Friedman, Emerson, Witzki, & Howerter, 2000; Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000; Salthouse, 2005, 2011; Salthouse et al., 2008). Individual differences in the ability to deal with complex intellectual tasks—where one must deal with various encoded mental representations and their relationships simultaneously while there is a high likelihood that one mental representation interferes with another—are central to fluid intelligence.

Fluid Intelligence Measurement

In modern factor analysis literature (McGrew, 2009; Schneider & McGrew, 2012) the broad fluid intelligence factor is defined by three narrow factors as inductive reasoning (I), general sequential reasoning (RG; deductive reasoning), and quantitative reasoning (RQ). Inductive reasoning tasks such as analogies, classifications, series, and geometric matrices, such as those that appear in Raven's Progressive Matrices (Raven, Raven, & Court, 1998), one of the most studied tests of fluid intelligence, are considered the markers of fluid reasoning. Sequential reasoning involves deductive logic tasks, with conclusions being inferred from premises, and quantitative reasoning involves induction and deduction tasks with numbers and mathematical operations.

Most intelligence batteries contain one or more fluid intelligence tests. For instance, the Woodcock-Johnson Battery (Woodcock, McGrew, & Mather, 2001) includes Concept Formation (I) and Analysis Synthesis (RG); the Wechsler Intelligence Scales (Wechsler Preschool and Primary Scale of Intelligence—Third Edition, WISC–IV, and WAIS–IV; Weiss, Keith, Zhu, & Chen, 2013) have Matrix Reasoning (I), Picture Concepts (I), and Figure Weights (RQ); the Stanford–Binet Intelligence Scale (SB5; Roid & Pomplun, 2012) includes Object-Series/Matrices (I), Early Reasoning, Verbal Analogies/Verbal Absurdities, and Verbal and Non-Verbal Quantitative Reasoning (RQ); the Differential Abilities Scales (DAS-II; Elliot, 2012) has Matrices (I), Pictures Similarities (I), and Sequential and Quantitative Reasoning (I and RQ); the Kaufman Assessment Battery for Children (KABC-II; Singer, Lichtenberger, Kaufman, Kaufman, & Kaufman, 2012) includes Conceptual Thinking (I), Pattern Reasoning (I), Story Completion (RG); and the Reynolds Intellectual Assessment Scales (Reynolds, Kamphaus, & Raines, 2012) includes Odd-Item Out (I) and Verbal Reasoning.

Given the diversity of tests already available, why develop a new measure? One reason is that all of the available measures were developed before the new cognitive and neuroscientific understanding of fluid intelligence was available. This research shed light on specific features of cognitive tasks that will engage the desired latent process underlying fluid intelligence. Essentially, whenever an intellectual task involves an increased amount of information that must be dealt with simultaneously and when there is a high likelihood that one mental representation interferes with the processing of another, core controlled attention executive processes of working memory will be engaged, revealing individual differences in fluid intelligence.

A fundamental definition of validity states that "a test is valid for measuring a theoretical attribute if and only if variation in the attribute causes variation in the measurement outcomes through the response process that the test elicits" (Borsboom & Mellenbergh, 2007). This conception has been incorporated in the latest version of the *Standards for Educational and Psychological Testing,* because it relates to evidence of test score interpretation based on the analysis of response process (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Therefore, if test items are built to inherit these features as part of the explanation of their difficulty, the test will potentially have more construct relevant variance built into it (Messick, 1980).

A second point is that the objective definition of cognitive complexity in terms of manipulable task features, called complexity factors, is a key element for developing generalized scales that transcend a particular test and understanding other tests of fluid intelligence. Complexity factors are broad-level sources of variance of item difficulty. In IRT, item difficulty is positioned on the same scale as a person's abilities. One advantage of a common metric for ability and item difficulty is that the explanations for the variability in item difficulty are also explanations for the variance of the ability scale (Andrich, 1990). This commonality is especially important because item variables that are supposed to relate to specific aspects of cognitive complexity can be easily manipulated in experimental studies that test the effects of these aspects on item difficulty (Embretson & Reise, 2000; Primi, 1995, 1998, 2002). Although item variables are easily manipulated, the same is not true for people. In these studies, "the relative difficulty of the various underlying cognitive process are operationalized by varying item stimulus features" (Embretson, 1998, p. 383). Hence, the explanation of item difficulty provides an operational definition of the ability scale. Linking item variables to item difficulty parameters augments the construct validity of test score interpretations, provided that these item variables are linked theoretically and empirically to different demands on the underlying cognitive processes (i.e., to cognitive complexity; Embretson, 1983, 1994). This linkage traces back to Carroll's (1993) concept of behavioral scaling, or the "process of stating in behavioral terms what tests results directly imply with regard to what examinees know or can perform" (p. 299). Construct validity is enhanced by making "reference to the level of cognitive processing involved in correct response at a given item difficulty level" (Carroll, 1993, p. 304).

Because complexity factors—linked to construct interpretations—are in a higher level of abstraction, they can be useful for analyzing several groups of intelligence tests that use the same stimuli. Such analyses make it possible to create scales that are test independent and thereby realize one of the important advantages of IRT. In summary, item difficulty can be predicted by complexity factors that, in turn, are more general and can be applied to different tests that measure the same construct. Thus, these variables establish a basis for constructing common scales to describe the level of cognitive processing that underlies items from similar tests. Ultimately, these variables could be used to link different tests on the basis of underlying task complexity.

A third point is that new tests, based on these modern methods, contribute to overcoming a problem referred to as arbitrary metrics, which is defined as the lack of information on "how a one-unit change on the observed score reflects the magnitude of change on the underlying dimension" (Blanton & Jaccard, 2006, p. 28). One consequence of this problem is the lack of a clear understanding of clinically significant change in psychological interventions. By having a scale with an objective metric of cognitive complexity mapped onto component processes that were mastered while solving the items, one can reach a less arbitrary metric for test score interpretation. Such an instrument would provide rich information for diagnosis and monitoring change that would be more useful for clinicians and researchers.

A fourth advantage of having a more objective scale of fluid intelligence is that practitioners and researchers would be stimulated to be aware of the core task features and complexity factors that are central to the definition of the latent scale formed by the items. An objective scale stimulates thinking about what cognitive process a person has mastered, what kinds of cognitive tasks will present difficulties when he or she is faced with similar demands, or what a specific change in test scores after an intervention would mean in terms of the underlying cognitive process. These considerations facilitate diagnostic practice, planning, and monitoring of interventions. An objective scale is also important for analyzing other tests. Even those that are not explicitly stated as measures of fluid intelligence could be tapping similar cognitive functions if their tasks vary in difficulty due to the complexity factors summarized in the present research. Finally, this research provides an updated test score interpretation for fluid intelligence tests, with reference to the underlying basic process of working memory and executive functions, based on more recent literature on cognitive psychology and neuropsychology.

In summary, although various tests of fluid intelligence are available, they do not possess an objective scale metric mapped on cognitive complexity and core process underlying fluid intelligence. The only test identified on the basis of these principles was the Abstract Reasoning Test (ART; Embretson, 1998), which was created with the purpose of illustrating these new methods. The goal in the present study was to illustrate the application of the cognitive design system for test development (Embretson, 1994, 1998) in the development of a scale for fluid intelligence that combines cognitive psychology and Rasch modeling. Although it was applied to a specific test presented in this article, this method is expected to contribute to intelligence testing in general because it demonstrates the development of criterion-referenced norms for test score interpretations for the underlying scale based on cognitive psychology and modern psychometric methods, in addition to norm-referenced measures of individual differences. An objective complexity scale is proposed that is general enough to represent any test of fluid intelligence, provided that the test's items are mapped onto the complexity scale presented in this article. This study contributes to the construct validity of a more general spectrum of interpretations of fluid intelligence test scores.

Method

Participants

The participants were 2,797 freshmen undergraduate students from the University of São Francisco, a private university in a small city in the state of São Paulo, Brazil. The ages ranged from 17 to 66 years ($M = 22.2$ years, $SD = 5.5$ years), and the distribution was positively skewed (skewness = 2.4); approximately 65.4% of the participants had ages in the range of 17–21 years. The group included 1,622 women (58%) and 1,175 men (42%). The participants came from 18 majors that required between 4 and 6 years of study; in some cases, their diploma is equivalent to a bachelor's degree. These majors represented nine general areas: medicine (4.4%), dental medicine (8.2%), psychology (17.2%), business (24.2%), engineering (9.9%), law (21.6%), education (4.2%), literature (2.9%), and tourism (7.4%).

Procedure

Data were collected for the present study as part of a basic abilities-assessment program for freshmen students that was financed by the university. The goal of the program was to investigate the reasoning and basic knowledge abilities of first-year students entering college. The program included the assessment of logical reasoning, reading comprehension, general literacy, moral judgment, and fluid intelligence. The participants were volunteers, and the testing was part of their regular academic activities. The participants completed the tests in two sections that lasted an average of 90 minutes each. After analysis of the data, the participants received a personal report that informed them of their scores and explained what the scores meant.

Test Development

A prototype task used for the measurement of fluid intelligence involves geometric inductive matrix items (Marshalek et al., 1983), such as Raven's Progressive Matrices, which is exemplified in the upper half of Figure 1. In earlier studies, Primi (1995, 1998, 2002) and Primi, Cruz, Muniz, and Petrini (2006) experimentally manipulated matrix item variables to investigate the impact of these features on item complexity. These earlier studies provided detailed examples of each item feature and an explanation of the link with fluid intelligence capacities. In summary, three main features were identified that were considered the sources of item complexity in geometric inductive matrix problems: (a) the amount of information due to the increase in the number of elements or in the number of rules

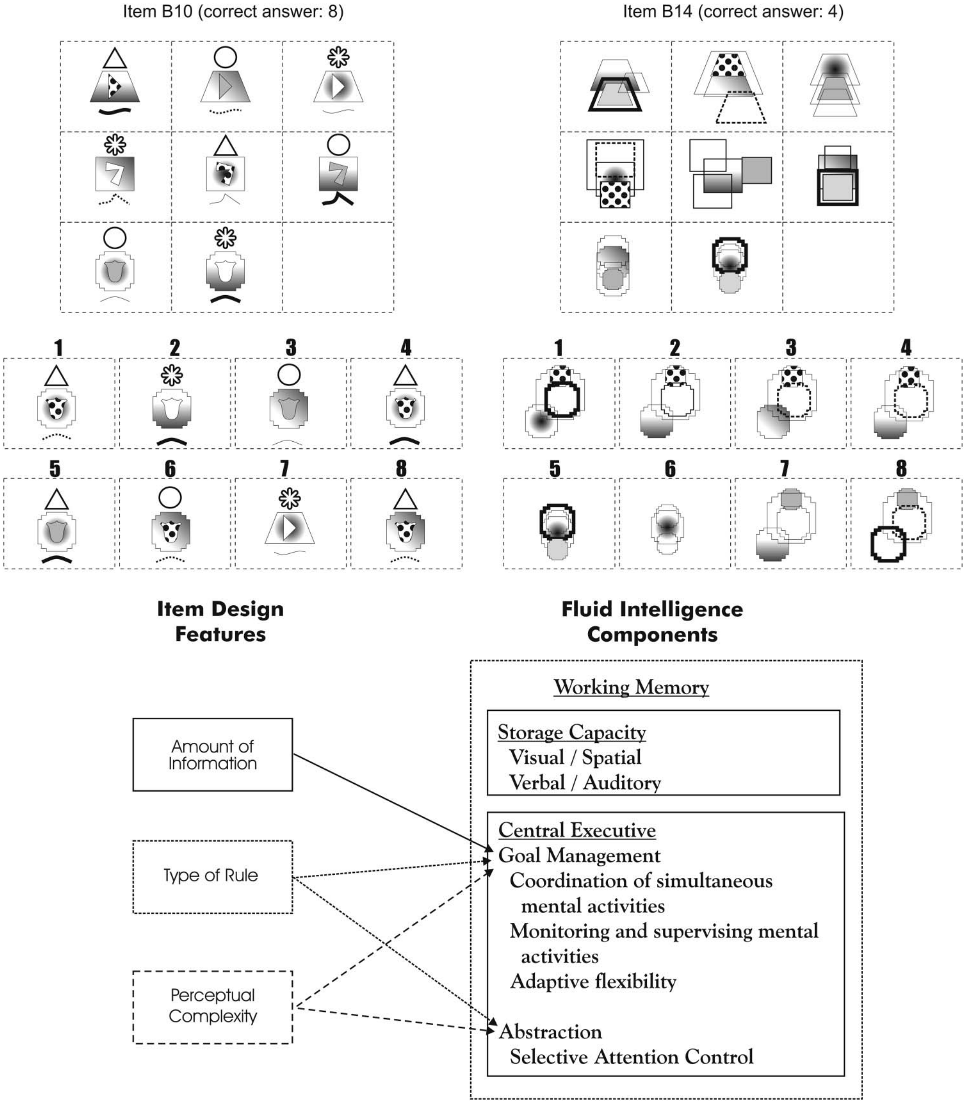

Figure 1. Examples of fluid intelligence items used in the present study and a summary of sources of complexity for geometric matrix items and their link with fluid intelligence capacities.

relating these figures, (b) the complexity of these rules, and (c) the perceptual complexity of the stimulus.

Amount of Information: Number of Elements and Rules

The number of elements refers to the number of geometric figures or attributes that exist in a specific matrix problem, whereas the number of rules refers to the number of relationships that exist among the different elements or attributes. Each matrix cell of Item B10 in Figure 1 is composed of four geometric figures (four elements), and, across the row, these figures change their properties, such as shape, outline, and fill. Increases in amount of information are supposed to interfere with the central executive, which stresses the need for processing information while storing the products of these processes and the eventual need for changing strategies when the problems become highly complex.

Type of Rules: Content of the Relationship

The complexity of the type of rules refers to the nature or content of the relationships or transformations that are applied to elements or attributes. A summary of these rules was presented in previous works (Primi, 1995, 1998, 2002). These rules are postulated to possess different levels of complexity due to the differential demands on controlled and selective attention (abstraction capacity). According to Carpenter et al. (1990), people induce rules in a serial manner, trying simpler rules before more complex ones. Therefore, more complex rules will be associated with a greater information load, because a person must consider several rules and remember the results of these earlier attempts before arriving at a complex rule.

Perceptual Complexity: Element Salience

The third feature is perceptual complexity, which refers to the likelihood of misleading cues that produce ambiguity. Perceptual complexity refers to the amplification or diminishment of the visual salience of the elements that must be grouped to solve the problem (Hornke, 2002; Meo, Roberts, & Marucci, 2007; Pérez-Salas, Streiner, & Roberts, 2012; Primi, 2002). For example, Item B14 is a perceptually complex version of Item B10 in which the elements are not aligned as in B10, which breaks the good continuity, and all of the geometric figures possess the same shape, which facilitates grouping by similarity. For perceptually complex items, the likelihood of forming irrelevant groups of elements based on perceptual features is increased compared to that for perceptually simpler items. This feature is therefore central in creating interference and intensifying the need of controlled and selective attention ability, because creating the correct grouping requires that certain fragments of the perceptual field must be selected and focused on while others must be ignored (abstraction). The increased likelihood of creating irrelevant groupings implies that more attempts to solve the problem also impose demands on the cognitive functions of coordination, monitoring, and rerouting.

Item Design Variables and Cognitive Process of Inductive Reasoning

The bottom part of Figure 1 presents the hypothetical theoretical relationships between the complexity factors and the cognitive process that underlie fluid intelligence. The processes summarized so far were taken from experimental psychology studies that manipulated the item features and tested their effects on item complexity. Different types of arrows have been extended from each factor to indicate the most probable components on which effects are likely to be observed.

The first thing to note from this diagram is that all of the complexity factors make demands on goal management, although the specific operations that they require appear to differ depending on the factor. The factor amount of information indicates the relevant extent of information overload on the working memory that must be considered. Once the limit of the storage capacity has been reached, some effective type of sequential management becomes necessary; this implies the use of hierarchical strategies to organize the flow of information in a limited space, thus involving the goal management system. The type of rule and perceptual complexity interfere at the moment when the elements or attributes are perceptually grouped (encoded), which requires the efficient use of selective attention. These variables can also overload the goal management system when the likelihood of forming irrelevant groupings increases, especially for more perceptually complex items. In the case of complex items, the storage capacity is surpassed because of the increase in the number of attempts. The role of goal management is the implementation and monitoring of these attempts and the operation of discarding the irrelevant attribute relationships that were encoded. According to this diagram, complexity factors should not have a specific and independent effect on a particular capacity; rather, all of the factors should influence all capacities. The only exceptions include the amount of information, which does not appear to have any effect on abstraction, and perceptual organization, which seems to affect abstraction in particular.

Test Design

Two paper-and-pencil tests were composed. Forms A and B each consisted of 16 items similar to the ones presented in Figure 1, all of which were selected from the 64-item pool studied by Primi (2002). This earlier study contained two computerized forms, each containing 32 items that were defined by the conditions of an experiment that manipulated the complexity factors: number of elements (2 or 4), number of rules (2 or 4), types of rules (four categories discussed below), and perceptual complexity (low or high). Thus, each form of 32 items corresponded to the factorial combinations of these variables ($2 \times 2 \times 4 \times 2 = 32$). For the present study, a subset of items from this earlier study was selected on the basis of its psychometric properties (dispersion of difficulty and item–total correlations). These items were used to compose two new paper-and-pencil tests to be used in the basic abilities assessment program, rather than the computerized versions used earlier. Six items appeared on both forms to function as anchor items. This equated the scores from Form A and Form B and resulted in an item bank of $10 + 10 + 6 = 26$ distinct items. Hence, each form consisted of 10 unique items and six anchor items.

Each item consisted of a 3×3 matrix with an empty cell and eight response alternatives. Each of the problems was composed of an organized set of geometric figures that obeyed either two or four rules; the participant was required to discover these rules so that he or she could generalize from them and select the most appropriate of the eight options offered to fit into the blank space. Two examples of the items used in the current study are presented in Figure 1.

The item structure was determined by the three variables: amount of information (number of elements and number of rules), type of rule, and perceptual complexity. An item consisted of either two or four geometric figures, termed elements, and those elements were related by either two or four rules. The rules employed to relate the items were (a) quantitative pairwise progressions (i.e., the increase or decrease of some attribute from one element to the next, such as size, shading, number series, or shape); (b) figure addition and subtraction, in which the figure in one cell was composed of a combination of two figures present in two other cells; (c) the distribution of three values, in which the elements were instances of a conceptual attribute, such as shading, inclination, color, size, outline, and shape; (d) attribute addition, in which an element was created by the combination of two attributes from the other two elements; and (e) the distribution of two values, in which element subparts appeared in only two of the three elements in the three cells of the matrix row.

Two levels of perceptual complexity were also used: easy to encode and difficult to encode. The manipulations employed to produce these two levels have been detailed elsewhere (Primi, 1995, 1998, 2002); in summary, the manipulations were based on gestalt principles of perceptual grouping, such as similarity and good continuation, to increase the likelihood of forming irrelevant groupings. For instance, by altering the relative positions of corresponding elements across a row, it is possible to increase the complexity involved in forming relevant groups due to the interruption of natural perceptual continuity that would have facilitated the grouping. These principles can also be used to create a perceptual tendency to group elements according to the appropriate conceptual rule (e.g., by harmonically aligning corresponding elements in space) and thus increase the formation of groups based on good continuity. Each easy-to-encode item was always accompanied by a difficult-to-encode complement that was consistent with all of the other structural variables; that is, the number of elements, number of rules, and type of rule involved.

The two examples in Figure 1 comprised four elements and four rules, each of which involved the distribution of three rules. The item on the left (B10) was easy-to-encode, and the item on the right was B10's difficult-to-encode congener.

The seven incorrect options were systematically created to be increasingly more correct. Hence, the options were created in such a way that there were one or two options from each of the following categories: (a) two alternatives with only one incorrect transformation, (b) two alternatives with two incorrect transformations, (c) two alternatives with more than two incorrect transformations, and (d) one alternative that was a copy of the cell to the left of the blank space.

Earlier studies have investigated the correlation of the test with external variables. The simple correlation was .62 with Raven's Advanced Progressive Matrices, .42 with general sequential reasoning, .46 with spatial visualization, and .36 with crystallized intelligence. A latent variable modeling of a general factor using Raven's Advanced Matrices indicated that the loadings from the Form A and B scales were .86 and .78, respectively (Prieto & Velasco, 2006; Primi, 1998, 2002; Primi, Santos, & Vendramini, 2002).

Data Analysis and Statistical Models

The data analysis was performed in four steps. First, the classical psychometric statistics (item difficulties, item–total correlations, and internal consistency reliabilities) were calculated. Second, a full-information item factor analysis was performed with TESTFACT software (Wilson, Wood, & Gibbons, 1991) to test the unidimensionality assumption (Hambleton, Swaminatham, & Rogers, 1991). This method analyzes response vectors instead of correlation matrices and uses all of the information available from the participant's responses, not just the covariance among items, as is the case in traditional methods of factor analysis (Embretson & Reise, 2000).

Third, the IRT parameters of Rasch's dichotomous and Rasch–Master's partial credit models (Wright & Masters, 1982; Wright & Stone, 1979) were calculated to calibrate the item difficulties and person measures by the use of unconditional maximum likelihood estimation performed by WINSTEPS (Linacre, 2011). The model fit was evaluated by two indices: information-weighted fit (*infit*) and outlier-sensitive fit (*outfit*). Infit and outfit are indices that are summary of the residuals "reported as mean-squares, chi-squared statistics divided by their degrees of freedom, so that they have a ratio-scale form with expectation 1 and range 0 to +infinity" (Linacre & Wright, 1994a, p. 350). For large samples, fit indices between 0.7 and 1.3 indicate a good fit. In general, fit indices in the range of 0.5–1.5 are considered productive for measurement, whereas values above 1.5 are considered unproductive for scale construction (Linacre & Wright, 1994b).

The final step involved a multiple regression analysis to predict item difficulty from item indicators of complexity. The dependent variable was the item difficulty parameter b_i, and the predictors were variables related to the amount of information, including a series of five dummy codes related to the type of rules (quantitative pairwise progressions, figure addition and subtraction, distribution of three values, attribute addition, and distribution of two values) and perceptual complexity. This final step consisted of mapping the item difficulty scale onto these objective variables of item features that were, in turn, linked to underlying cognitive processes. In this phase, the variables were tested for their capacity to predict item difficulties.

Results

Classical Psychometric Properties

Table 1 shows the descriptive statistics for the total score. Each item was scored either as 0 or as 1, and the total score varied from 0 to 16. The table also shows the frequency distribution of item difficulty and the biserial correlations between the items and the total scores. The item difficulties varied from .29 to .82, and the majority of the biserial correlations were above .60, thus contributing to a respectable internal consistency reliability of .80 and .82. These results indicate that the goals pursued were achieved; that is, the two forms represented a wide range on the complexity continuum and simultaneously presented good psychometric qualities. The original test forms from which the items were selected con-

Table 1
Descriptive Statistics for Total and Item Scores of Test Forms A and B

Statistic	Form A		Form B	
Descriptive statistics for total scores				
M	8.5		9.4	
SD	3.9		3.7	
α	.82		.80	
Descriptive statistics for items	*p*	r_b	*p*	r_b
M	.54	.67	.59	.66
SD	.16	.07	.17	.09
Min	.30	.51	.29	.49
Max	.80	.80	.82	.79
Frequency distributions[a]	*f* (%)	*f* (%)	*f* (%)	*f* (%)
<.31			1 (6.3)	
.31–.40	4 (25.0)		3 (18.8)	
.41–.50	1 (6.3)		1 (6.3)	1 (6.3)
.51–.60	6 (37.5)	2 (12.5)	1 (6.3)	4 (25)
.61–.70	2 (12.5)	9 (56.3)	4 (25)	6 (37.5)
.71–.80	2 (12.5)	5 (31.3)	5 (31.3)	5 (31.3)
>.81	1 (6.3)		1 (6.3)	

[a] Difficulty index (*p*) and biserial correlations between item scores and total scores for r_b.

stituted 32 items each and had test score internal consistency coefficients of .84 and .85. Thus, despite the reduction in the number of items, these new test forms were nearly as reliable as the original, longer forms (Primi, 2002).

Item Factor Analysis

A separate full-information item factor analysis was performed by TESTFACT for each test form using the data from 1,339 participant for Form A and 1,308 participant for Form B. For Form A, the mean of tetrachoric correlation for 120 item pairs was .47; the eigenvalues for the first and second factors were, respectively, 8.09 and .99, and all of the item loadings on the first unrotated factor were above .45. These results clearly indicate the existence of a dominant factor that was responsible for the covariance among the items. A model that contained two factors did not provide a significant increase in the model fit ($\Delta\chi^2 = 13.5$, $df = 15$, $p > .50$). The results for Form B were slightly different than those for Form A; the mean of tetrachoric correlation for 120 item pairs was .42; the eigenvalues for the first and second factors were, respectively, 7.44 and 1.05, and all of the item loadings on the first unrotated factor were above .48. However, unlike for Form A, a small but a significant increase in the model fit was found for Form B when a second factor was extracted ($\Delta\chi^2 = 42.45$, $df = 15$, $p < .001$). The correlation between the first and the second factor with an oblique promax rotation of factor loadings was .82, which indicates that, even for Form B, it is still appropriate to assume a unidimensional solution.

The full-information factor analysis results provided evidence that the item sets were measuring the same construct. The unexpected results for Form B led to a further investigation of the factor loading pattern in the two-factor solution, and the easy-to-encode items were found to be separated from the difficult-to-encode ones. This finding suggests that the two types of items may represent specific separate, albeit highly correlated, components of the construct being assessed. More evidence that these factors do not violate unidimensionality comes from a second check conducted via a principal components analysis of linearized residuals performed by WINSTEPS. The variance of the second factor, in eigenvalue units, was 1.29 and 1.27 for Forms A and B, respectively. This accounted for nearly 8% of the total variance, which was much less than the 45% to 50% of the variance that was accounted for by the first dominant factor found in the full-information item factor analysis. These values were below 1.4, which has been indicated in simulation studies by Raiche (2005) to be the minimum magnitude for considering a residual component as a dimension that is differentiated from random noise.

Rasch and Partial Credit Analysis

Given the evidence that a single underlying factor is responsible for the performance on the items, the next step was to locate each item on the factor continuum (i.e., to estimate both the item difficulty and the person's location on the same continuum). The analysis for this section was performed with WINSTEPS software (Linacre, 2011). First, item and person parameters were examined, centering on the items, and then the infit and outfit indexes were examined to identify problems with model fit. Table 2 presents the estimates of item difficulty and their standard errors (*b* for Rasch and b_1 and b_2 for partial credit, which will be explained next) and the infit and outfit indices.

For Form A, the infit indexes varied from .85 to 1.19 ($M = .99$, $SD = .08$) and the outfit statistics varied from .71 to 1.37 ($M = 1.04$, $SD = .18$); 3 of the 16 items had an outfit above 1.2. For Form B, the infit indexes varied from .88 to 1.14 ($M = .99$, $SD = .07$) and the outfit statistics varied from .80 to 1.28 ($M = 1.01$, $SD = .15$); 3 of the 16 items had an outfit above 1.2. The fit indices revealed a good fit to the model. Only one item (Form A: 16) showed an outfit greater than 1.3.

The item difficulties for Form A varied from −1.67 to 1.41 ($M = 0$ and $SD = .96$), and the root-mean-square error of approximation (*RMSE*) was .07. The resulting separation index, $G_I = 13.67$, is "a ratio scale index comparing the 'true' spread of the measures with their measurement error. It indicates the measure of spread of this sample of examinees (or test items) in units of the test error in their measures" (Fisher, 1992, p. 238). For Form B, the item difficulties varied from −1.45 to 1.75 ($M = 0$ and $SD = .99$), and the *RMSE* was .07, resulting in a separation index of 14.10. The Rasch item separation reliability for both forms was high (.99). In conclusion, considering that item difficulty generally varies from −3 to 3, the difficulty indexes were somewhat restricted in range, but they were still quite distinct, stable, and replicable. This finding is important because the relative difficulty order among the items is expected to be stable and clearly separated, thus defining a direction for the underlying variable.

The second aspect investigated in this step was the application of the Rasch–Masters partial credit model (Wright & Masters, 1982). This test was composed so that each of the seven incorrect options was located at a different distance from the correct answer. This placement permitted the assignment of partial scores for these

Table 2
Rasch Difficulty and Partial Credit Threshold Estimates With Indexes of Misfit for Items of Forms A and B

Form and item	Rasch				Partial credit					
	b	*SE*	*Infit*	*Outfit*	b_1	*SE*	b_2	*SE*	*Infit*	*Outfit*
Form A										
Anchor items										
(4) 44sph	0.00	0.06	1.06	1.11	−0.44	0.08	0.23	0.06	0.97	0.97
(5) 22sph	−1.44	0.07	0.85	0.71	−0.60	0.11	−1.10	0.07	0.81	0.71
(7) 24sph	−0.72	0.07	0.97	0.97	−1.51	0.12	−0.23	0.06	0.84	0.81
(9) 22spd	−0.80	0.07	0.98	0.95	−0.32	0.09	−0.58	0.07	0.98	0.94
(15) 44spd	1.12	0.07	1.02	1.29	−0.46	0.08	1.29	0.07	1.04	1.05
(16) 24spd	1.41	0.07	1.02	1.37	−0.04	0.07	1.47	0.07	1.10	1.24
Unique items										
(1) 22cph	−1.10	0.07	0.96	0.93	0.09	0.09	−1.02	0.07	0.94	0.88
(2) 22coh	−1.67	0.08	0.86	0.78	0.85	0.09	−2.01	0.08	0.87	0.82
(3) 24coh	−0.28	0.06	1.01	1.01	0.34	0.07	−0.32	0.06	1.06	1.07
(6) 44coh	−0.24	0.06	0.94	0.96	−0.90	0.09	0.10	0.06	0.87	0.84
(8) 44cph	0.20	0.06	0.95	0.94	0.71	0.07	−0.01	0.06	1.02	1.04
(10) 22cpd	−0.07	0.06	1.04	1.06	0.42	0.07	−0.15	0.07	1.12	1.26
(11) 22cod	0.62	0.06	0.97	0.95	−0.73	0.08	0.87	0.06	0.95	0.94
(12) 24cpd	1.41	0.07	1.03	1.08	0.75	0.06	1.20	0.07	1.05	1.08
(13) 44cpd	1.31	0.07	1.02	1.20	0.58	0.07	1.17	0.07	1.09	1.17
(14) 44cod	0.24	0.06	1.19	1.30	0.03	0.08	0.31	0.06	1.17	1.23
Form B										
Anchor items										
(5) 44sph	−0.32	0.07	0.97	0.89	−0.21	0.09	−0.05	0.06	0.97	0.89
(4) 22sph	−0.86	0.07	0.90	0.80	−0.33	0.10	−0.55	0.07	0.93	0.91
(7) 24sph	−0.40	0.07	1.14	1.24	−1.51	0.13	0.12	0.06	1.01	1.06
(9) 22spd	−0.51	0.07	1.01	0.99	0.00	0.09	−0.30	0.07	1.04	1.04
(11) 44spd	1.42	0.07	1.07	1.14	−0.68	0.08	1.74	0.07	1.04	1.03
(16) 24spd	1.75	0.07	0.99	1.15	0.19	0.07	1.91	0.07	1.05	1.07
Unique items										
(1) 42coh	−0.57	0.07	1.03	1.21	−0.59	−1.80	−0.20	0.07	1.04	1.19
(2) 42cph	−1.28	0.08	0.88	0.83	−1.07	−0.68	−0.77	0.00	0.85	0.76
(3) 22cph	−1.45	0.08	0.90	0.83	−0.68	−1.30			0.85	0.75
(6) 24cph	−0.58	0.07	0.97	0.96	−0.60	−0.72	−0.21	0.07	0.95	0.92
(8) 44cph	−0.44	0.07	0.95	0.93	0.44	−1.04	−0.40	0.07	1.07	1.10
(10) 44coh	−0.69	0.07	0.98	0.92	−0.05	−0.80	−0.46	0.06	0.98	0.87
(12) 24cpd	0.45	0.06	0.98	0.98	−0.31	−0.70	0.74	0.06	1.00	1.00
(13) 42cod	1.21	0.07	1.13	1.28	−0.42	0.00	1.50	0.07	1.08	1.13
(14) 44cod	1.28	0.07	1.02	1.03	0.59	−0.15	1.31	0.07	0.96	0.96
(15) 22cpd	0.98	0.07	1.00	1.04	0.51	−1.80	1.03	0.07	1.16	1.19

alternatives. In the analysis above, all seven options were considered incorrect, but some of them could have been considered partially correct. The question of whether assigning a partial score would increase reliability remained to be addressed.

For this analysis, each item was scored from 0 to 2. The score 2 was given for correct answers, the score 1 was given for partially correct answers, and the score zero was given for completely incorrect answers. For items that involved four rules, partial credit was given for the alternatives with one, two, or three correct rules; for items that involved two rules, partial credit was given for the alternatives with one correct rule. Although items could have more than one partially correct answer, not all of the answers functioned in the same way or provided the same psychometric information. Another approach is to provide information if it is possible to observe a distinct region in the ability continuum in which its probability of occurrence is greater than that of the alternative. If this pattern is not observed, there is no useful information in the alternative.

Plots of the empirical curves that displayed the probability of each alternative as a function of ability were examined to investigate which alternatives were worth partial scoring. These plots were obtained from the multiple choice distractor analysis procedure provided by the RUMM 2010 software (Andrich, Lyne, Sheridan, & Luo, 1997). The curves of each possible partially correct alternative were analyzed for a pattern that indicated the existence of useful psychometric information. The presence of such a pattern resulted in an alternative that was effectively eligible for partial scoring.

After this definition, a new data matrix was prepared for each form that contained partial scorings. The partial credit analysis developed by Wright and Masters (1982) was applied with WINSTEPS software. The right side of Table 2 shows the results of the analysis. Table 2 presents, for each item, the threshold estimates b_1, which indicates the point on the logit scale at which the likelihood of scoring 1 becomes greater than scoring 0, and b_2, which indicates the point at which the likelihood of scoring 2 becomes greater than scoring 1. Table 2 also shows the standard errors of the threshold estimates and the misfit indexes.

For Form A, the item infit indices varied from .81 to 1.17 (M = .99, SD = .10), the outfit indices varied from .71 to 1.26 (M = 1, SD = .16), and 3 of the 16 items showed an outfit greater than 1.2. The *RMSE* was .04, and the item separation index was 13.9. For Form B, the item infit indices varied from .85 to 1.16 (M = 1, SD = .08), and the outfit indices varied from .75 to 1.19 (M = .99, SD = .13). The *RMSE* was .05, and the item separation index was 11.7. For both forms, the item separation reliability was high (.99).

The use of the partial credit model improved the reliability of individual person measures from .74 to .77. These results were lower than the classical reliabilities reported in Table 1 because of differences in the way that classical and Rasch reliabilities are calculated (Linacre, 1997). Feldt's (1980) test for the comparison of two dependent reliability coefficients showed that the increases in reliability obtained from the partial credit model were statistically significant (for Form A, t = 10.61, df = 1399, p < .001; for Form B, t = 9.49, df = 1394, p < .001).

The upper portion of Figure 2 shows the scatter diagram positioning for students according to their Rasch logit and partial credit logit scores. Before this graph was plotted, the theta measures on the two systems and their standard errors were linearly equated via the procedures proposed by Wright and Masters (1982).

Although the correlation between the two test scores was high (r = .97, N = 2,797, p < .001) for a given Rasch logit, the partial credit logit scores were more widely spread, especially at the low end of the ability continuum. Due to this spread, the use of partial scores contributed to a greater differentiation between individuals. To test whether the students located at the low end were more dispersed than the students located at the high end when partial credit scores were used, the unstandardized residuals were obtained from a regression analysis of the effects of the scores from the Rasch logit scores on scores from the partial credit logit. The participants were then split into two groups: those with a Rasch logit score below the 16th percentile (low-end group, n = 511), and those with a Rasch logit score above 84th percentile (high-end group, n = 570). The residuals of these groups were examined. For the low-end group, the mean of the unstandardized residuals was .10, and for the high-end group, the mean of the unstandardized residuals was .11; however, the standard deviations were .51 and .36, respectively. The Levene's test of equality of variances was significant (F = 13.46, p < .001), which indicates that the residuals from the low-end group were more dispersed than the residuals from the high-end group. These results indicate that the prediction of the partial credit logit scores from the Rasch logit scores was less effective at the low end of the ability scale because the partial credit provided more information for the differentiation of these students than for the differentiation of the students at the high end of the scale. The correlation between the Rasch logit and partial credit logit scores was .81 (n = 511, p < .001) for the low-end group and .95 (n = 570, p < .001) for the high-end group.

Another method to evaluate the advantage of partial credit scoring is to examine the information function. The information of a test is the inverse of the square root of the measurement error, and this information varies across ability levels (Hambleton et al., 1991). For this comparison, the information was transformed into local test score reliabilities, as suggested by Daniel (1999), because this reliability metric is more meaningful than the original metric that was obtained from the information function. The lower half of Figure 2 portrays the local test score reliabilities as a function of ability and shows the means of these reliabilities for 15 groups of people. The total scores ranged from 1 to 15, as obtained from the Rasch analysis (dotted gray line) and the partial credit analysis (solid black line). It is evident that partial credit scoring produced a great increase in test score reliability for the individuals located at the lower end of the ability continuum. At the high end, the two methods were almost equally efficient, although there was a slight decrease in the efficiency of the partial credit scoring method.

Construct Representation: Building a Fluid Intelligence Scale

Once the prerequisite conditions were met, the final and most important step was conducted to investigate whether the item design features were able to predict item difficulty. As summarized in Figure 1, three design features were thought to have an effect on item complexity due to their demands on the cognitive components of fluid intelligence: the amount of information, the type of rule, and perceptual complexity.

The goal in the present analysis was to predict the Rasch difficulty index on the basis of these independent variables using multiple regression. All of the variables were entered into the equation to investigate their unique effects. Four variables reached the criteria of p < .10: perceptual organization, the amount of information, and figure addition and subtraction. The structural variables resulted in a multiple correlation coefficient of R = .930 and R^2 = .865, which was statistically significant, F(18, 7) = 16.50, p < .0001. Table 3 presents the final parameters that were obtained from the multiple regression. The table also shows the matrix of correlations between the predictors and the difficulty parameter. As previously reported, these results indicate that perceptual organization had the strongest impact, increasing item complexity by 1.482 logits. The amount of information was slightly less important than perceptual organization, increasing item complexity by .172 logits.

These results show a good fit between the cognitive model and the observed item complexities. The three design variables, particularly the first two, captured most of the variance in item complexity, leaving only a small portion of the variance unexplained. Therefore, the item complexity and ability continuum can be translated into objective quantifications of task complexity factors that, in turn, can be linked to cognitive components of fluid intelligence.[1] Such linkage can potentially contribute to the creation of criterion-referenced individual test scores that describe which cognitive operations a person has demonstrated depending on his or her measures on the fluid intelligence scale.

[1] An example of the integration and a common scale of ability estimates, normative references, item examples, and objective item complexity features can be found in http://www.labape.com.br/gf/gf.html

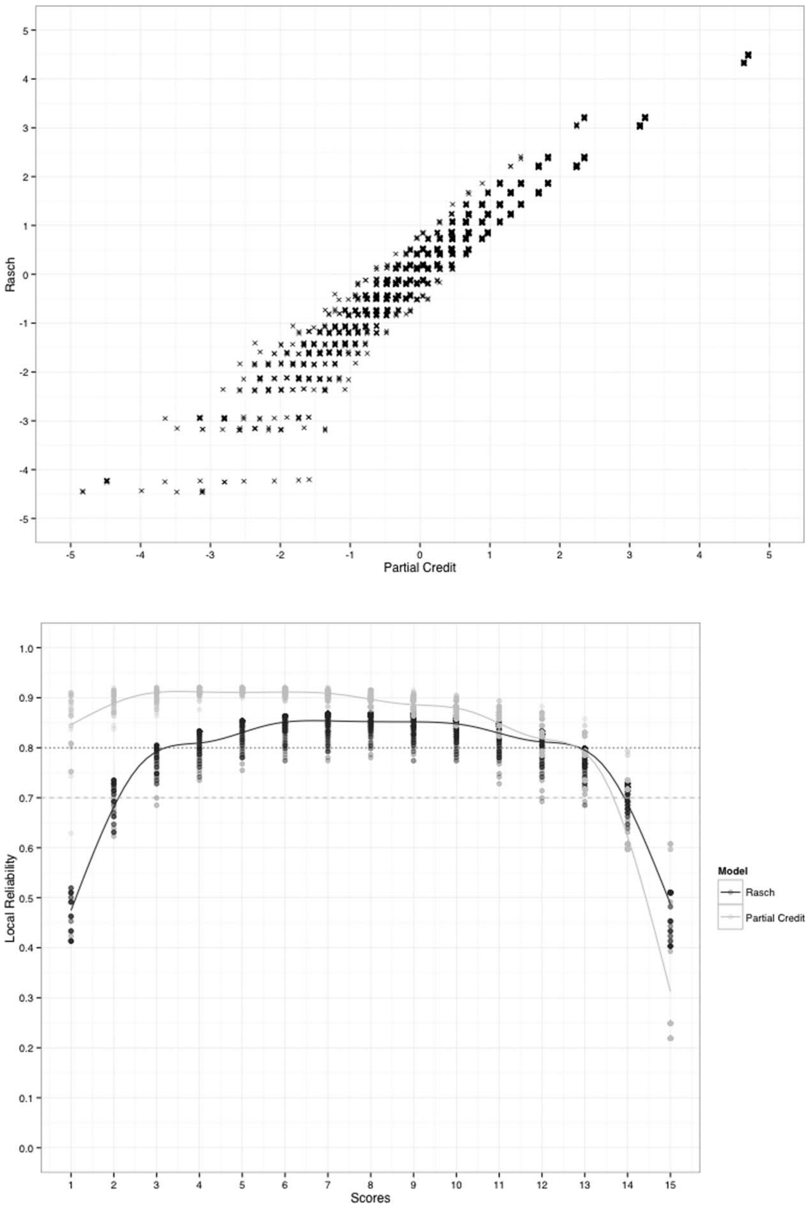

Figure 2. Scatter diagram of Rasch versus partial credit logits and local reliability for 16 levels of the ability continuum with Rasch model and partial credit model.

Discussion

Cognitive testing has been criticized because it has not yet satisfactorily incorporated new knowledge available from cognitive psychology and psychometrics into test construction. This study contributes to the literature by applying Embretson's (1994) method of the development of a fluid intelligence scale. Instead of focusing on the development of a test, this study focused on the definition of a cognitive complexity scale for fluid intelligence linking objective information about task difficulty for the purpose of fomenting criterion-referenced test score interpretations based on concepts obtained from cognitive psychological research. There are three major implications of this study: One relates to the practical utility of IRT models in test development, the second relates to the construct representation of fluid intelligence, and the third concerns the integration of cognitive psychology and psychometrics in intelligence test interpretation.

Psychometric Implications

The first series of psychometric analyses has indicated that these items are quite robust for defining a variable; even with a small number of items, the two tests displayed similar or better psychometric properties than other existing tests of fluid intelligence, such as the old Raven's Advanced Progressive Matri-

Table 3
Summary of Results of Regression Analysis Predicting Item Difficulty From Structural Variables

Item design (structural) variables	*B*	*SE*	β	*t*	*p*
Intercept	−1.836	.544		−3.375	.003
Perceptual organization	1.482	.171	.763	8.673	.000
Amount of information	.172	.081	.295	2.135	.047
Quantitative pairwise progressions	−.489	.386	−.213	−1.266	.222
Figure addition and subtraction	.837	.466	.312	1.797	.089
Distribution of 3 values	−.412	.241	−.189	−1.710	.105
Attribute addition	.327	.312	.142	1.045	.310
Distribution of 2 values	.309	.239	.141	1.292	.213

Correlation matrix between all variables in the regression

	1	2	3	4	5	6	7
1. *b* (item difficulty parameter)	—						
2. Perceptual organization	.787***	—					
3. Amount of information	.393*	−.032	—				
4. Quantitative pairwise progressions	.038	.042	.008	—			
5. Figure addition and subtraction	.241	.033	.306†	.778***	—		
6. Distribution of 3 values	−.146	−.040	−.060	−.332*	−.259	—	
7. Attribute addition	.229	.042	.394*	−.300†	−.234	.491**	—
8. Distribution of 2 values	.153	.134	−.269†	−.332*	−.259	.218	.285†

Note. For each item, the amount of information was indexed as the number of geometric figures (1 point for 2 elements and 2 points for 4 elements) plus the number of rules involved (1 point for 2 rules and 2 points for 4 rules) plus one point if two rules were applied to a single element. The possible values for the amount of information were 2, 3, and 4. An extra point was given if more than one rule was mixed in the same geometric figure, because the rules must be mentally separated to be processed; this requires more mental resources than if they were separated over two different geometric figures. The type of rule was represented by nine dummy-coded variables, each indicating whether a particular type of rule was present in the item (quantitative pairwise progression, figure addition, distribution of three values, attribute addition and distribution of two values). These variables were not summed, because they represented a qualitative attribute that indexed the content of the rules that were related to the geometric figures. The presence of a rule did not necessarily indicate more complexity because some of the rules were simpler than others. Perceptual organization was represented by a dummy code that indicated the presence or absence of a perceptually complex configuration (0 or 1).
† $p < .10$ and $p > .05$. * $p < .05$. ** $p < .01$. *** $p < .001$.

ces (Raven et al., 1998) and the new ART (Embretson, 1998). Thus, this article shows that knowledge from cognitive psychology about the determinants of item complexity is a useful tool in guiding item production that results in an item set that possesses good psychometric qualities. Additionally, the items in the current set possessed better construct representation than items from other studies, because the theory that was used in the current study to predict item difficulty was based on the cognitive processes that underlie fluid intelligence (Cornoldi, 2006; Deary, 2001; Embretson, 2002, 2006; Hornke, 2002). Thus, this article presents a method for developing items that have a solid grounding in cognitive psychology and that are tuned to the construct that they are supposed to measure. Additionally, this article links the item difficulty scale to objective item structural variables. This linkage has an interesting psychometric implication for equating different tests. If these item variables become useful for predicting the item difficulties of the different tests that assess fluid intelligence, they can be used as superordinate common items with which to equate the ability estimates from different tests. Thus, item complexity theory consists of a method that changes the focus from development of tests to a more general development of objective variables that provide more fundamental scales for a construct. Of particular importance was the demonstration that test score reliability can be increased without lengthening the test if a partial credit model is applied. The model used in this study contained a systematic scoring technique that was based on a careful evaluation of the incorrect options.

Implications for Construct Representation of Fluid Intelligence

The second step in the analysis investigated the use of item features in the prediction of item complexity. The results were similar to those obtained in a previous study; however, the magnitude was greater in the current study (Primi, 2002). Whereas in the previous study, perceptual complexity and the amount of information accounted for .642 of the variance in item difficulty, in the present study, these two variables, combined with a variable that was associated with the type of rule, accounted for .865 of the variance in item complexity. This is a noticeable difference, and it is well above the amount that has been documented in the literature. For instance, Embretson (1998) accounted for .773 of the variance and Hornke (2002) accounted for .40 of the variance. The literature has demonstrated a consistent impact of these variables on item complexity. This impact has important implications for the construct representation of fluid intelligence. The effect of perceptual organization increased the item complexity by approximately 1.5 logits. Only the people who score above the median are proficient in solving this type of item. What makes difficult-to-encode items complex, and what cognitive processes underlie item solving?

It has become increasingly evident that fluid intelligence is linked to the process of executive controlled attention of working memory (Conway et al., 2002; Engle et al., 1999; Kane, 2005 Miyake et al., 2000; Oberauer et al., 2000; Salthouse, 2005, 2011; Salthouse et al., 2008). A fundamental feature of tasks that demand controlled attention is the existence of a condition of interference; that is, the presence of conflicting information that competes for attentional focus (Kane et al., 2001). Difficult-to-encode items meet this condition because of two features: the break of good continuity among the geometric figures and the presence of irrelevant similarities that make the constituent elements less salient (Pérez-Salas et al., 2012). The first feature makes the geometric figures more mixed and provides interference. For instance, when a person must form two groups of related figures, in the moment at which he or she analyses one group to induce the relationships among the figures in that group, the other group of figures will be in the same visual field. The presence of the second group potentially interferes with the focus of attention, thus demanding central executive resources for active maintenance of the relevant figures and for blocking the irrelevant figures. The second feature makes the creation of an incorrect group of figures more likely and thus demands central executive resources for blocking these activated representations and for managing the additional processing required to find the correct solution. In contrast, easy-to-encode items have groups of figures that are clearly separated in the visual field. These figures are more visually salient and possess markedly different attributes, such as shape and colors, which diminishes the demands made on the central executive.

It is interesting to note the greater magnitude of the effect of perceptual complexity compared with the effect of the amount of information. This difference suggests that an increase in interference and irrelevant information has a greater debilitating effect on performance than does the increase in relevant information. Based on these results, it is possible to hypothesize that the deactivation of inferred relationships that have proved to be irrelevant, which is linked to a basic process of inhibition, consumes many more resources from the central executive than does the processing of relevant information. In summary, these results indicate that these variables affect item complexity because they increase the amount of controlled and effortful processing requirements that make up the central feature of fluid intelligence tasks, as demonstrated by Salthouse et al. (2008).

Intelligence Test Score Interpretation: Integration Norm and Criterion-Referenced Interpretations

One important contribution for test construction is the linking between cognitive processing comprehension of fluid intelligence and objective item features aligned with the theta scale, which can further interpretation of criterion-referenced test scores based on the precepts of cognitive psychology and neuropsychology. Traditionally, the results of intelligence tests are interpreted with a normative reference that compares a person with the expectations of a normative group. This approach provides information concerning the location of a person relative to a group of similar people but does not by itself provide much information about a person's performance capacity—a problem that has been criticized under the name of arbitrary metrics (Blanton & Jaccard, 2006; Embretson, 2006). The lack of information about capacity stems in part from a lack of studies that have evaluated the cognitive processes involved in intelligence tasks and that would support validity evidence for response process (Borsboom & Mellenbergh, 2007). This article is an attempt to develop a scale that integrates cognitive psychology in the production of a criterion-referenced interpretation of what a person is able to perform and the nature of the person's ability in terms of item complexity factors mapped onto cognitive components of the task response process, for various levels of the scale. Only a few other studies have attempted this type of application for cognitive tests, including a study on behavior scaling by Carroll (1993) and one on item-referenced meaning by Embretson and Reise (2000). For instance, an average ability level is characterized by capability to coordinate and monitor simultaneous processing and storage activities to organize a logical sequence of steps and infer up to four relationships simultaneously, provided that the elements are organized in such a way that the likelihood of irrelevant relationship formation is reduced. An above average ability (upper quartile) is characterized by capability to use the same process but with more than four steps and with less organized information, which will eventually lead to relationships that would have to be discarded. Upper quartile ability also indicates capability of making an analytical decomposition of visual patterns and a reorganization of more abstract visual concepts possible, as well as ability to maintain this abstract representation in face of interference. This type of test score interpretation is less arbitrary and, when integrated with normative and criterion-related information, can be very helpful; for instance, in diagnosis and in appreciating clinically significant changes. Imagine an intellectually high-functioning person who has a neurological condition that has impaired his fluid capacity in such a way that his score decreased from a normative superior quartile to an average score. With such information, it would be possible to better describe his impairment.

In future studies assessing fluid intelligence, the scale developed here can be employed to provide an interpretation for test scores in terms of cognitive processes, thus contributing to the understanding of cognitive psychological functioning in relation to psychometric measures. Considering the recent literature linking working memory, executive functions, and fluid intelligence, it is possible to reconsider fluid intelligence tests as important measures of executive functions in a neuropsychological battery. The links among test scores, task parameters, and cognitive process associated to executive function can improve the understanding of test scores by providing more substantial information on the construct being assessed. This could be particularly useful when fluid intelligence tests are used in neuropsychological diagnosis and monitoring of interventions.

It is also expected that new tests of cognitive abilities will gradually integrate the methodologies discussed here, providing test users with criterion-referenced measures in addition to norm-referenced information. These methodologies will make possible a deeper understanding of human intelligence by integrating the advances of cognitive psychology with intelligence testing toward new cognitive diagnostic models.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131,* 30–60. doi:10.1037/0033-2909.131.1.30

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, D. (1990). Ability of an item. *Rasch Measurement Transactions, 4*(2), 101. Retrieved from http://www.rasch.org/

Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (1997). *RUMM2010 Rasch unidimensional measurement models* [Computer software]. Duncraig, Australia: RUMM Laboratory Pty.

Baddeley, A. D., & Hitch, G. J. (1994). Developments in the concept of working memory. *Neuropsychology, 8,* 485–493. doi:10.1037/0894-4105.8.4.485

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence, 8,* 205–238. doi:10.1016/0160-2896(84)90009-6

Blair, C. (2006). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. *Behavioral and Brain Sciences, 29,* 109–125.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist, 61,* 27–41. doi:10.1037/0003-066X.61.1.27

Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 85–115). Cambridge, United Kingdom: Cambridge University Press.

Burgess, G. C., Gray, J. R., Conway, A. R. A., & Braver, T. S. (2011). Neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span. *Journal of Experimental Psychology: General, 140,* 674–692. doi:10.1037/a0024695

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review, 97,* 404–431. doi:10.1037/0033-295X.97.3.404

Carroll, J. B. (1993). Test theory and the behavioral scaling of test performance. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 297–322). Hillsdale, NJ: Erlbaum.

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30,* 163–183. doi:10.1016/S0160-2896(01)00096-4

Cornoldi, C. (2006). The contribution of cognitive psychology to the study of human intelligence. *European Journal of Cognitive Psychology, 18,* 1–17. doi:10.1080/09541440500215889

Daniel, M. H. (1999). Behind the scenes: Using new measurement methods on the DAS and KAIT. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 37–63). Mahwah, NJ: Erlbaum.

Deary, I. J. (2001). Human intelligence differences: Towards a combined experimental–differential approach. *Trends in Cognitive Sciences, 5,* 164–170. doi:10.1016/S1364-6613(00)01623-5

Duncan, J., Burgess, P., & Emslie, H. (1995). Fluid intelligence after frontal lobe lesions. *Neuropsychologia, 33,* 261–268. doi:10.1016/0028-3932(94)00124-8

Duncan, J., Emslie, H., & Williams, P. (1996). Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive Psychology, 30,* 257–303. doi:10.1006/cogp.1996.0008

Duncan, J., Schramm, M., Thompson, R., & Dumontheil, I. (2012). Task rules, working memory, and fluid intelligence. *Psychonomic Bulletin & Review, 19,* 864–870. doi:10.3758/s13423-012-0225-y

Elliot, C. D. (2012). The Differential Ability Scales–Second Edition. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 336–357). New York, NY: Guilford Press.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197. doi:10.1037/0033-2909.93.1.179

Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107–135). New York, NY: Plenum Press.

Embretson, S. E. (1995). The role of working memory capacity and general control process in intelligence. *Intelligence, 20,* 169–189. doi:10.1016/0160-2896(95)90031-4

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380–396. doi:10.1037/1082-989X.3.3.380

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 219–250). Mahwah, NJ: Erlbaum.

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist, 61,* 50–55. doi:10.1037/0003-066X.61.1.50

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128,* 309–331. doi:10.1037/0096-3445.128.3.309

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika, 45,* 99–105. doi:10.1007/BF02293600

Fisher, W., Jr. (1992). Reliability statistics. *Rasch Measurement Transactions, 6*(3), 238. Retrieved from http://www.rasch.org/

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence, 24,* 79–132. doi:10.1016/S0160-2896(97)90014-3

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience, 6,* 316–322. doi:10.1038/nn1014

Hambleton, H. K., Swaminatham, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Heitz, R. P., Unsworth, N., & Engle, R. W. (2005). Working memory capacity, attention control, and fluid intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 61–78). Thousand Oaks, CA: Sage.

Hornke, L. F. (2002). Item-generation models for higher order cognitive functions. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 159–178). Mahwah, NJ: Erlbaum.

Kane, M. J. (2005). Full frontal fluidity? Looking in on the neuroimaging of reasoning and intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 141–163). Thousand Oaks, CA: Sage.

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General, 130,* 169–183. doi:10.1037/0096-3445.130.2.169

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131,* 66–71. doi:10.1037/0033-2909.131.1.66

Kyllonen, P. C., & Christal, R. (1990). Reasoning ability is (little more than) working memory capacity?! *Intelligence, 14,* 389–433. doi:10.1016/S0160-2896(05)80012-1

Linacre, J. M. (1997). KR-20 or Rasch reliability: Which tells the “truth”? *Rasch Measurement Transactions, 11*(3), 580. Retrieved from http://www.rasch.org/

Linacre, J. M. (2011). Winsteps (Version 3.72.3) [Computer software]. Retrieved from http://www.winsteps.com/

Linacre, J. M., & Wright, B. D. (1994a). Chi-square fit statistics. *Rasch Measurement Transactions, 8*(2), 350. Retrieved from http://www.rasch.org/

Linacre, J. M., & Wright, B. D. (1994b). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(2), 370. Retrieved from http://www.rasch.org/

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence, 7,* 107–127. doi:10.1016/0160-2896(83)90023-5

McGrew, K. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research, *Intelligence, 37,* 1–10. doi:10.1016/j.intell.2008.08.004

McGrew, K. S., & Wendling, B. J. (2010). Cattell–Horn–Carroll cognitive-achievement relations: What we have learned from the past 20 years of research. *Psychology in the Schools, 47,* 651–675. doi:10.1002/pits.20497

Meo, M., Roberts, M. J., & Marucci, F. S. (2007). Element salience as a predictor of item difficulty for Raven’s Progressive Matrices. *Intelligence, 35,* 359–368. doi:10.1016/j.intell.2006.10.001

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012–1027. doi:10.1037/0003-066X.35.11.1012

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howerter, A. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology, 41,* 49–100. doi:10.1006/cogp.1999.0734

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology, 12,* 252–284. doi:10.1016/0010-0285(80)90011-0

Oberauer, K., Süß, H. M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity: Facets of a cognitive ability construct. *Personality and Individual Differences, 29,* 1017–1045. doi:10.1016/S0191-8869(99)00251-2

Pérez-Salas, C. P., Streiner, D. L., & Roberts, M. J. (2012). A comparison between element salience versus context as item difficulty factors in Raven’s Matrices. *Intelligence, 40,* 325–332. doi:10.1016/j.intell.2012.04.001

Prabhakaran, V., Smith, J. A. L., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. E. (1997). Neural substances of fluid reasoning: An fMRI study of neocortical activation during performance of the Raven’s Progressive Matrices test. *Cognitive Psychology, 33,* 43–63. doi:10.1006/cogp.1997.0659

Prieto, G., & Velasco, A. D. (2006). Visualização espacial, raciocínio indutivo e rendimento acadêmico em desenho técnico [Spatial visualization, inductive reasoning and academic achievement in technical drawings]. *Psicologia Escolar e Educacional, 10*(1), 11–20. doi:10.1590/S1413-85572006000100002

Primi, R. (1995). *Inteligência, processamento de informação e teoria da gestalt: Um estudo experimental* [Intelligence, information processing and gestalt theorie: An experimental study]. Unpublished master’s thesis, Pontifical Catholic University of Campinas, Campinas, Brazil.

Primi, R. (1998). *Desenvolvimento de um instrumento informatizado para avaliação do raciocínio analítico* [Development of a computerized test for the assessment of the analytical thinking]. Unpublished doctoral dissertation, University of São Paulo, São Paulo, Brazil.

Primi, R. (2002). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence, 30,* 41–70. doi:10.1016/S0160-2896(01)00067-8

Primi, R., Cruz, M. B. Z., Muniz, M., & Petrini, M. C. (2006). Validade de construto de um instrumento informatizado de avaliação dinâmica da inteligência fluida [Construct validity of a computerized instrument for the dynamic assessment of fluid intelligence]. *Psico, 37,* 109–122.

Primi, R., Ferrão, M. E., & Almeida, L. S. (2010). Fluid intelligence as a predictor of learning: A longitudinal multilevel approach applied to math. *Learning and Individual Differences, 20,* 446–451. doi:10.1016/j.lindif.2010.05.001

Primi, R., Santos, A. A., & Vendramini, C. M. M. (2002). Habilidades básicas e desempenho acadêmico em ingressantes universitários [Basic abilities and academic achievement in freshmen undergraduate students]. *Estudos de Psicologia, 7,* 47–55. doi:10.1590/S1413-294X2002000100006

Raiche, G. (2005). Critical eigenvalue sizes (variances) in standardized residual principal components analysis (PCA). *Rasch Measurement Transactions, 19*(1), 1011–1012. Retrieved from http://www.rasch.org/

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven’s Progressive Matrices and Vocabulary Scales: Section 1. General overview*. Oxford, United Kingdom: Oxford Psychologists Press.

Reynolds, C. R., Kamphaus, R. W., & Raines, T. C. (2012). The Reynolds Intellectual Assessment Scales (RIAS) and the Reynolds Intellectual Screening Test (RIST). In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 400–421). New York, NY: Guilford Press.

Roid, G. H., & Pomplun, M. (2012). The Stanford–Binet Intelligence Scales, Fifth Edition (SB5). In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 249–268). New York, NY: Guilford Press.

Salthouse, T. A. (2005). Relations between cognitive abilities and measures of executive functioning. *Neuropsychology, 19,* 532–545. doi:10.1037/0894-4105.19.4.532

Salthouse, T. A. (2011). What cognitive abilities are involved in trail-making performance? *Intelligence, 39,* 222–232. doi:10.1016/j.intell.2011.03.001

Salthouse, T. A., Pink, J. E., & Tucker-Drob, E. M. (2008). Contextual analysis of fluid intelligence. *Intelligence, 36,* 464–486. doi:10.1016/j.intell.2007.10.003

Schneider, W. J., & McGrew, K. (2012). The Cattell–Horn–Carroll model of intelligence. In D. Flanagan & P. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 99–144). New York, NY: Guilford Press.

Shimamura, A. P. (2000). Toward a cognitive neuroscience of metacognition. *Consciousness and Cognition, 9,* 313–323. doi:10.1006/ccog.2000.0450

Silvia, P. J., & Beaty, R. E. (2012). Making creative metaphors: The importance of fluid intelligence for creative thought. *Intelligence, 40,* 343–351. doi:10.1016/j.intell.2012.02.005

Singer, J. K., Lichtenberger, E. O., Kaufman, J. S., Kaufman, A. S., & Kaufman, N. L. (2012). The Kaufman Assessment Battery for Children—Second Edition and The Kaufman Test of Educational Achievement—Second Edition. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 269–296). New York, NY: Guilford Press.

Sternberg, R. J. (1977). A component process in analogical reasoning. *Psychological Review, 84,* 353–378. doi:10.1037/0033-295X.84.4.353

Sternberg, R. J. (1978). Isolating the components of intelligence. *Intelligence, 2,* 117–128. doi:10.1016/0160-2896(78)90003-X

Sternberg, R. J. (1981). Testing and cognitive psychology. *American Psychologist, 36,* 1181–1189. doi:10.1037/0003-066X.36.10.1181

Sternberg, R. J., & Gardner, M. K. (1983). Unities in inductive reasoning. *Journal of Experimental Psychology: General*, *112*, 80–116. doi:10.1037/0096-3445.112.1.80

Tatsuoka, K. K. (2009). *Cognitively diagnostic assessment: An introduction to the rule space method*. New York, NY: Taylor & Francis.

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013). WAIS–IV and clinical validation of the four-and five-factor interpretative approaches. *Journal of Psychoeducational Assessment*, *31*, 94–113. doi:10.1177/0734282913478030

Whitely, S. E. (1980). Modeling aptitude test validity from cognitive components. *Journal of Educational Psychology*, *72*, 750–769. doi:10.1037/0022-0663.72.6.750

Whitely, S. E., & Schneider, L. M. (1981). Information structure for geometric analogies: A test theory approach. *Applied Psychological Measurement*, *5*, 383–397. doi:10.1177/014662168100500312

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis*. Chicago, IL: Scientific Software International.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. London, United Kingdom: Erlbaum.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III tests of cognitive abilities*. Itasca, IL: Riverside.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: Mesa Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: Mesa Press.

Received November 1, 2012
Revision received March 10, 2014
Accepted March 18, 2014 ■