Distinguishing Verbal, Quantitative, and Nonverbal Facets of Fluid Intelligence in Young Students

Joni M. Lakin

Auburn University

James L. Gambrell

University of Iowa

Keywords: Fluid reasoning; School children; Bifactor models

Author Note

Joni M. Lakin, Ph.D., Department of Educational Foundations, Leadership, and Technology, Auburn University. James L. Gambrell, Psychological and Quantitative Foundations, University of Iowa

The author gratefully acknowledges the helpful comments of David Lohman on earlier drafts of this article.

Correspondence concerning this article should be addressed to Joni Lakin, Department of Educational Foundations, Leadership, and Technology, Auburn University, Auburn, AL 36831. Email: joni.lakin@auburn.edu

Abstract

Measures of broad fluid abilities including verbal, quantitative, and figural reasoning are commonly used in the K-12 school context for a variety of purposes. However, it is difficult to differentiate these domains for young children (grades K-2) who lack language and mathematics literacy. This study evaluated the latent factor structure of a picture-based test of broad fluid reasoning abilities using a bifactor model to separate general and broad domain factors in a large representative sample of U.S. school children. Results showed substantial evidence that picture-based item formats can distinguish between general and domain-specific fluid reasoning abilities. The verbal tasks showed the strongest domain factor and discriminant validity, although the quantitative tasks also showed considerable strong evidence of a domain factor. Results indicate that distinct verbal and quantitative reasoning domains can be uncovered in the early school grades.

Distinguishing Verbal, Quantitative, and Nonverbal Facets of Fluid Intelligence in Young Students

Measures of broad fluid reasoning abilities including verbal, quantitative, and figural reasoning are commonly used in K-12 school contexts for a variety of purposes including educational placement, instructional differentiation, and ability-achievement discrepancy assessment (Corno, 1995; Gregory, 2004). Measuring separate verbal and quantitative domains is essential for purposes that require identifying strengths and weaknesses by school subject. The critical challenge in the assessment of these abilities is the measurement of uniquely verbal and quantitative fluid reasoning in young children (grades K-2) who lack basic language and mathematics literacy skills. The present study evaluated the ability of picture-based item formats to measure distinguishable content factors on a reasoning test. Specifically, we explored the extent to which three broad domain factors were recovered in addition to the overall fluid reasoning factor.

Measuring distinct verbal, quantitative, and nonverbal/figural domains is relatively straightforward for students who are literate in both the language of instruction and mathematical symbol systems (i.e., the majority of students in grade 3 and above). However, verbal and quantitative abilities are more difficult to measure in younger students due to their lack of literacy. For young students, many group-administered tests resort to teacher-read item prompts which can create significant demands on students' receptive language skills, thereby introducing construct-irrelevant variance and reducing discriminant validity between domains. Such demands on receptive language (e.g., with English as a common language between student and teacher) make it especially difficult to fairly and validly assess students who are culturally and linguistically diverse (Lohman & Gambrell, 2012).

Although there is some debate on the differentiation of reasoning abilities in young children (Juan-Espinosa, García, Colom, & Abad, 2000; Kane & Brand, 2006; Keith & Reynolds, 2010), researchers have consistently found the presence of strong broad ability factors (beyond *g* or Gf) in all school-aged students (Carroll, 1993; Kane & Brand, 2006). Even in mathematics, which might be expected to depend on schooling and to develop later than other skills, research has shown that quantitative reasoning skills begin to develop before children are taught to count and before they are exposed to formal mathematics education (Starkey, 1992).

Assessing Fluid Reasoning across Content Domains

Fluid reasoning (Gf) can be defined as the process of drawing defensible inferences from incomplete information and it is a central component of cognitive abilities (Carroll, 1993; Gustafsson, 1984; Schneider & McGrew, 2012). Reasoning ability can vary by domain (i.e., the *content* being reasoned about), leading individuals to reason more effectively in some domains than others. The division of fluid reasoning processes by content or domain is supported by extensive empirical research (Beauducel, Brocke, & Liepmann, 2001; Carroll, 1993; Wilhelm, 2005; Lohman, 2000). In Carroll's (1993) compendium of factor analytic studies, three subfactors of fluid reasoning were identified: sequential reasoning, inductive reasoning, and quantitative reasoning. Wilhelm (2005) argues that these reasoning factors may be better understood as content factors rather than process factors with verbal, figural, and numerical content factors defining Gf. This is consistent with Carroll's findings (based on the typical tasks identifying each factor) and is consistent with faceted models of intelligence such as the Berlin Model of Intelligence Structure (BIS) which hypothesize both a content dimension (verbal, numerical, and figural) and a process dimension (reasoning, knowledge, and memory) to intelligence tests (Beauducel et al., 2001; Suβ & Beauducel, 2005; Wilhelm, 2005). Beauducel et

al. (2001) argue that measuring a content facet in addition to a process facet results in greater construct validity due to aggregating the fluid reasoning process across content. In addition to the benefits of aggregation, measuring the three subfactors of Gf also allows the test to align with the reasoning demands of the typical classroom which includes considerable verbal and quantitative content as well as demands on general abstract reasoning (Corno et al., 2002; Snow & Lohman, 1984; Wilhelm, 2005).

Purpose of the Study

In this study, we were interested in exploring the construct validity of picture-based measures of verbal and quantitative reasoning for young students (grades K-2). The picture-based item formats examined come from the Cognitive Abilities Test Form 7 (CogAT 7; Lohman, 2011). CogAT is a multidimensional (and multi-level) ability test developed for grades K-12 which has a long history of use in schools and well-regarded psychometric properties (DiPerna, 2005; Gregory, 2004). CogAT is one of the most widely used group ability tests in both the United States and the United Kingdom (where a parallel form of CogAT is used, abbreviated CAT). The test consists of three batteries that measure verbal, quantitative, and figural reasoning with three item formats in each battery. In previous editions, the test levels designed for grades K-2 consisted of verbal and quantitative items that relied on teacher-read oral prompts with picture-based response options (e.g. CogAT 6; Lohman & Hagen, 2001). This yielded highly correlated verbal and quantitative batteries (Lohman & Hagen, 2002), which was inconsistent with the behavior of the grade 3-12 test levels (where quantitative and nonverbal/figural correlated more strongly).

CogAT 7 introduced picture-based item formats for young students that are analogous to verbal and quantitative formats used at the higher grade levels. These picture-based formats were

designed to draw on conceptual (verbal) reasoning and rudimentary quantitative reasoning. These formats are described in greater detail in the methods section. An added bonus of these formats is the potential to improve the fair and accurate assessment of these abilities for culturally and linguistically diverse students. Because the formats assume little shared language between teacher and student (apart from basic directions), the picture-based items are expected to reduce cultural and linguistic loading. As a result, we expected that smaller differences will be found between racial and ethnic groups, particularly Latino/Hispanic groups who are more likely to be current or former English-language learners. The authors describe the process used to select "culturally decentered" item content in the Research Guide (Lohman, in press b).

The CogAT 7 picture-based item formats are somewhat similar to the formats used by other tests, including the Universal Nonverbal Intelligence Test (UNIT; Bracken & McCallum, 1998) and the Kaufman Brief Intelligence Test (K-BIT2; Kaufman & Kaufman, 2004), which both use matrices formats with pictures of objects. However, the picture-based matrix items on these tests contain a mixture of conceptual and visual relationships between the pictures. Visual relationships (color, size, pattern differences) are most likely to draw on general and figural reasoning rather than specifically verbal reasoning. Only one other test was located that used a pictorial quantitative reasoning format, the Picture Sequence subtest of the cTONI. This test appeared to primarily measure quantitative concepts. However, the test is the only quantitatively oriented task in the battery and contributes to the pictorial rather than a quantitative composite. Thus, in contrast to existing tests, the picture-based formats studied here were designed to require students to identify distinctly conceptual (verbal) and quantitative relationships between the objects represented in the pictures. Through item selection, the tests also diminish the importance of visual features in item solutions.

Though the picture formats are similar to those on the tests above, only CogAT 7 attempts to use them to measure Gf content factors and thus provides the recommend minimum of three indicators for each group factor and three group factors for Gf (Bollen, 1989; Carroll, 1997). Heterogeneity of item content yields a measure of Gf with high "referent generality" (Gustafsson, 2002; Coan, 1964) and less construct underrepresentation (Messick, 1989). This plurality of measures is helpful for assessing the validity of the new formats in terms of detecting distinct verbal and quantitative factors. The following validity questions guided this study:

1. Do the picture-based item formats measure specific verbal, quantitative, and nonverbal/figural reasoning factors in addition to general reasoning ability? In other words, do they show internal discriminant validity?
    a. Allowing for a strong general reasoning factor, we expect that an appropriate factor analysis will detect the presence of significant secondary "domain" factor loadings in the new verbal and quantitative item formats. Domain factor loadings for the figural formats are expected as well, but are less important to its validity.
2. Do the three battery-level scores formed from picture-based formats correlate strongly with measures of academic achievement? That is, do they show external discriminant validity across academic domains?
    a. We expect the (Picture) Verbal Battery to correlate more strongly with reading-related achievement scores than mathematics-related scores. For the Quantitative and Nonverbal Battery, we expect the opposite.
3. Do scores derived from picture-based items reduce group mean differences between racial/ethnic, socioeconomic, or language proficiency groups compared to scores derived from similar tests using teacher-read oral prompts?

a. We expected to find smaller subgroup mean differences on CogAT 7 compared to historical data on CogAT 6.

Method

Participants

Data were taken from the 2010 joint national standardization of the Cognitive Abilities Test Form 7 and the Iowa Assessments Form E (ITBS-E). CogAT 7 data on the picture-based formats was available for 18,153 students in grades K-2. Matched ITBS-E data was available for 6,320 of these students. For analyses involving only CogAT scores, the full sample was used. Basic demographic information on the samples is shown in Table 1. Students' status as English-language learners[1] (ELL) or eligible for free or reduced price lunch (FRL) were reported by the schools. Classifications were based on local policies. In this table, the ELL and FRL categories are also broken down by ethnicity. The sample was representative of the U.S. school population on most dimensions.

Table 1 about here

Measures

CogAT 7. The picture-based formats on CogAT 7 are used for levels 5/6, 7 and 8. These levels are designed for students in kindergarten through second grade. At each level the test is divided into three batteries: Verbal, Quantitative, and Nonverbal (Figural). Each battery consists of three subtests, for a total of nine tests at each level. The number of items at each level is shown in Table 2. Items overlap across levels, so the total number of unique items on each subtest is 32 (16 for Number Puzzles). This table also displays the reliability of each test,

[1] The term English-language learner is used to designate students who enroll in a school where English is the language of instruction, but who speak a language (or languages) other than English natively and do not have full English proficiency at school entry.

computed using Cronbach's alpha. Number Puzzle items at level 8 were omitted because the format switches from pictures to numbers at this level.

Table 2 about here

CogAT 7 provides four battery scale scores: Verbal, Quantitative, Nonverbal, and Alt-Verbal. The Alt-Verbal score is intended for use with ELL students and omits the Sentence Completion test (which is the only subtest that still requires receptive language in either English or Spanish). This provides a verbal score that is free of items requiring English proficiency. The battery scales were developed using 2PL IRT models to vertically equate test levels. While some attention will be given to these battery scores, the primary goal of the present study is to analyze the validity of each picture-based item format, rather than the CogAT 7 scales. For this reason, the raw item-level scores are the primary unit of analysis. For some analyses, the Standard Age Score (SAS) scale is used as well. These scores are on an IQ-like scale with M=100 and SD=16.

The nine CogAT 7 item formats are shown in Figure 1. All are adaptations of the formats designed for older children. The Sentence Completion task is the only format that is not wholly picture-based, because it requires comprehension of an oral prompt. We retain the test in our analysis to provide a known marker of specific verbal reasoning ability. The Picture Analogies and Picture Classification tasks are analogous to the Verbal Analogies and Verbal Classification formats used in grades 3 and above except that they rely on pictures rather than on words. They are intended to test a student's ability to reason with everyday real-world objects and concepts.

Figure 1 about here; to be printed in black-and-white

The Number Analogies format is similar to the Picture and Figure Analogies, but emphasizes basic mathematic computation and reasoning in the item solutions. Both relative quantities (more/less) and simple computation (all quantities below 10) are represented in the

items. The traditional Number Series format was adapted for use with younger children by introducing an abacus-like toy where students count the number of beads in each column and choose the option that continues the numerical series (e.g., 1, 2, 3, 1, 2, ?). Finally, the Number Puzzles format is an adaptation of a new equation balancing format being used at upper levels of CogAT 7 where students must identify the value represented by one or more geometric shapes. At grades K-2 students are required to select the picture that equalizes the number of objects carried by two trains.

The Nonverbal (figural) battery required minimal adaptation to be appropriate for students in grades K-2. In fact, the Figure Analogies and Figure Classification formats were already part of the primary battery for CogAT6. Adapting the Paper Folding format from the multilevel edition simply required items of appropriate challenge for younger students.

The picture-based formats evaluated in this paper were developed for the Cognitive Abilities Test Form 7 borrowing cultural de-centering methods advocated by cross-cultural assessment guidelines (Hambleton, Merenda, & Spielberger, 2005) Through extensive review by individuals from a variety of cultural backgrounds and experiences, only concepts (and pictorial representations of those concepts) that were rated as culturally neutral for all U.S. school children were used in the final assessment (Lohman, in press b).

Iowa Assessments Form E. The Iowa Assessments Form E (formerly known as the Iowa Tests of Basic Skills) was used to measure achievement in five domains: Vocabulary, Reading, Listening, Math, and Math Computation (Dunbar, Welch, Hoover, & Frisbie, 2011b). Scale scores were used for each of the five areas. Matched data was only available for grades 1 and 2. Vocabulary and Reading domains are assessed by having students read individual words and short sentences. The Listening test requires sustained listening skills as well as literal and

inferential comprehension skills at all three levels. The math skills tested in the level 7 and 8 Math battery include recognizing numbers, counting, representations of fractions, geometric figures and patterns, and more complex representations of numbers and place value as well as basic algebraic concepts. All levels of the math subtest use teacher-read oral prompts. Math Computation scores measuring addition and subtraction skills were only available for students in grade 2.

Procedure

Tests were administered in English to students by their regular classroom teachers or a local testing coordinator. The median test date was November 9th, 2010. All testing was completed between October 18th, 2010 and November 29th, 2010.

Design

Bifactor model. The first research question concerned whether there was evidence of secondary battery-specific domain factors (verbal, quantitative, and figural) in addition to a primary factor. This question was critical because previous tests using picture-based formats have not strongly distinguished between domain-specific reasoning and general reasoning abilities. To estimate the degree of verbal or quantitative loading on each picture item, a two-dimensional, two-parameter item response theory model (2PL MIRT; Gibbons & Hedeker, 1992; Rijmen, 2009) was used to estimate the bifactor model shown in Figure 2. The model is comprised of a single primary dimension for general fluid reasoning ability (Gf) and secondary Verbal, Quantitative, and Figural content dimensions.

Figure 2 about here

Bifactor models have a long history in cognitive testing (Holzinger & Swineford, 1937; Thomson, 1948; Jöreskog, 1969; Yung, Thissen, & McLeod, 1999), and belong to a class of

models variously referred to as “hierarchical” or “nested factor” models (Gustafsson & Aberg-Bengtsson, 2010). The defining characteristics of such models are 1) all covariances between factors are fixed at zero and 2) all factor loadings are at the first order and thus all factors are directly linked to observable indicators (no higher order factors). In a bifactor model, each item is assumed to measure two factors: a single general dimension common to all items and a more specific orthogonal (uncorrelated) dimension shared only by related items. These less general “secondary” factors may be substantive group or specific factors, or they may simply be nuisance factors or what Cattell & Tsujioka (1964) referred to as “bloated specifics”.

The more widely used second-order factor model is in fact a special case of the bifactor model (Chen, West, & Sousa, 2006; Gustafsson, & Balke, 1993), and the two approaches are often regarded as more or less equivalent. However, the bifactor approach has a number of practical and statistical advantages. First, the bifactor model is uniquely suited to dealing with the situation commonly encountered in construct measurement where a scale appears to be both unidimensional and multidimensional depending on how it is analyzed and who interprets the results (Reise, Moore, & Haviland, 2010). Second, because the bifactor model allows every item to load directly on the general factor it is easier to avoid identifying group factors that have no construct validity (Chen et al., 2006). Third, the independent predictive contributions of each group factor to other variables can be studied more easily than in a second order factor model. Finally, the bifactor approach makes it simple to decompose item and scale variance into percent contributions from each factor, which are easier to interpret than correlated factor loadings or item discriminations (Gustafsson & Aberg-Bengtsson, 2010).

The model was estimated on the K-2 CogAT-only sample using Bayesian MCMC estimation in Mplus 6.1 (Muthén, & Muthén, 2010). Rather than separating analyses by grade

level, the full K-2 sample was used in order to maximize both the number of items per subtest as well as the latent factor variance.

MCMC was used because it is well suited to dealing with the large number of students, items, and factors being estimated, as well as the large amount of missing data due to the cross-grade design. Four MCMC chains were run and the Gelman-Rubin diagnostic was used to assess convergence (Brooks & Gelman, 1998; Muthén, & Muthén, 2010). High autocorrelation was observed during estimation so results were based on every 10^{th} iteration. Model fit was evaluated using Bayesian posterior predictive checking (Rubin, 1984; Asparouhov & Muthén, 2010). To help evaluate the robustness of the resulting parameter estimates, the model was also run separately on independent halves of the sample so that the consistency of estimates from each half could be examined. When tests of significance were used, the alpha level was set at a conservative .001 due to the many tests being performed. For each item format, the percent of items with a significant domain loading and the average loading on each factor were calculated.

The model did not include subtest factors due to the additional computational complexity. As a result, an index of generality was needed to ensure that domain factors were not merely subtest-specific factors arising from a single strong format. Therefore, the percentage of significant item loadings across subtests was used to index the generality of each factor and ensure that it was not merely a subtest or "bloated specific" factor (Cattell, & Tsujioka, 1964).

External validity. The analyses above evaluate whether the picture formats load on factors orthogonal to both the Gf factor and the other domain factors. To further clarify what is measured by the broad factors, external discriminant validity was assessed by regressing the latent factors onto the observed Iowa Vocabulary, Reading, Listening, and Math scale scores.

Regression coefficients were estimated in Mplus by adding the Iowa scale scores and regression paths to the bifactor measurement model.

Correlations were also computed between CogAT 7 subtest scores and ITBS scale scores. To be consistent with the cross-grade approach used in the factor analysis, these correlations were computed across grade as well. The CogAT 7 as published does not provide subtest scores that are comparable across test levels, so vertically equated IRT scale scores for each subtest were computed using Winsteps 3.63 (Linacre, 2006). These scale scores were only used to compute correlations and cross-grade reliabilities. Reliability estimates were based on the average IRT standard error and the cross-grade standard deviations.

Group differences. As discussed in the introduction, one of the purposes of the picture-based item formats was to reduce group mean differences caused by oral language load. This hypothesis was investigated using observed and marginal mean SAS score (M=100, SD=16) differences between various focal groups.

Marginal mean differences estimate the independent effect of group membership after controlling for other background variables. This analysis was carried out using a linear mixed model (hierarchical linear model) controlling for the effects of ethnicity, gender, FRL, ELL, age, grade, school SES (defined as Title 1 eligibility), region of the country, and public vs. private school. The multilevel nature of the dataset was modeled using random intercepts for schools. This set of analyses was computed in SPSS 20.

Five focal groups were analyzed: English-language learners (ELLs), students on free or reduced-price lunch (FRL), and students with Hispanic, Black, or Asian ethnic backgrounds. Each focal group was compared to an appropriate comparison group (non-ELL, non-FRL, and White, respectively). To provide a comparison to similar tests using teacher-read oral prompts,

group differences on the three batteries from the CogAT 6 standardization sample collected in 2000 are contrasted with the results from the 2010-2011 CogAT 7 standardization data.

Results

Table 2 shows raw score descriptive statistics for the full CogAT 7 sample and the smaller sample who took both CogAT 7 and the Iowa Assessments. Table 3 shows the intercorrelations of the nine subtests across grades K-2. These intercorrelations indicate strong relationships between formats within batteries, although there are clear methods effects for items that use the analogies (3 subtests) or classification format (2 subtests).

Table 3 about here

Bifactor MIRT model

The four MCMC chains were run for a total of 10,000 iterations. Convergence criteria were met after 3,000 iterations and every 10^{th} iteration from the last 5,000 was used to compute results. This provides a buffer of 2,000 iterations post-convergence. The overall Bayesian posterior predictive p-value (PPP) was significant, but PPP values for all but a few items were non-significant. Further inspection showed that while statistically significant, deviations from observed values were very small in magnitude.

Table 4 summarizes the results of the bifactor model estimation. Results are presented for all items and for the subset of items showing significant domain loadings. The domain percent column shows the fraction of common variance accounted for by the domain factor. This controls for differences in error variance across the formats and approximates the expected size of the domain variance component on a sufficiently long scale composed of each item type.

Table 4 about here

Turning first to the verbal formats, 88% of the Picture Analogies, 94% of the Picture Classifications, and 88% of Sentence Completion items had significant domain loadings. On the quantitative formats, the loadings for Number Puzzles and Number Series were also consistent, with 100% and 69%, respectively, showing loadings on a secondary domain factor. Number Analogies showed weaker evidence of a domain factor with only 56% loading on a secondary factor. As expected, the figural formats showed less consistent domain measurement that was mostly confined to strong loadings from Paper Folding items.

Significant factor loadings do not necessarily correspond to strong factor loadings, however, so we next considered the percent of variance attributable to the domain-specific factor versus the general (Gf) factor. Verbal factor variance was similar for Picture Analogies and Sentence Completion at around 25%, with Picture Classification at 15%. Gf variance for Sentence Completion was the weakest out of all 9 formats (see column 3 in Table 4). Among the quantitative tests, Number Puzzles stood out as the strongest format, with the largest average variance on both Gf and domain factors out of all 9 formats. Number Series and Number Analogies offered similar average domain variance components when considering only items with positive domain loadings. However, Number Analogies had a substantial number of items (10 out of 32) with significant negative loadings. Further investigation showed 9 out of 10 of these items to be the easiest in the set, the first 9 items at level 5/6. Inspection of these items revealed them to be more easily solved by visual matching strategies.

The figural formats showed the least evidence of a strong domain factor although Gf loadings were high. Both Figure Matrices and Figure Classifications had less than half of the items loading on the domain factor and little domain variance even among items with significant

loadings. Only Paper Folding had domain variance comparable to what was seen for other domains.

The results reported above are based on the full sample of 18,153 students. To evaluate the replicability of these parameter estimates, the sample was randomly split into two halves and the model was fit to each half. Pearson correlations between general factor loading, domain factor loading, and item difficulty estimates across the two random samples were .99, .96, and .99 respectively, indicating strong replicability of the model results.

Prediction of Achievement

Regression coefficients between the latent factors and Iowa scales are shown in table 5. Results showed substantial predictive validity for the General and Verbal factors across the Iowa scales, but weaker validity for the Quantitative and Figural factors. On average, coefficients were around .66 for Gf, .32 for Verbal, .10 for Quant, and .05 for Figural. All relationships were significant except Vocabulary and Reading on Figural. Disattenuated correlations between CogAT and Iowa Assessment scores are presented in Table 6. Correlations ranged between .47 and .86 for subtests and .55-.86 for battery scale scores. The more verbal Iowa scales (Vocabulary, Reading, and Language) had an average correlation of r = .69 with CogAT 7 Verbal subtests and r = .58 for Quantitative and Figural subtests. For math tests, Verbal, Quantitative, and Figural subtests all had an average correlation of r = .61, indicating little discrimination. Looking at Math Computation alone (only available for grade 2), the average correlation was r = .49 with CogAT 7 Verbal subtests, r = .59 for Quantitative, and r = .55 for Figural subtests. Battery-level correlations ranged from .55-.86 and were generally highest for Verbal.

Table 5 about here

Table 6 about here

Group Differences

The changes made to CogAT 7 appear to have had a significant effect on group differences.[2] See Table 7. The largest changes were observed on the Verbal Battery where gaps between Hispanic, Black, ELL, and FRL students and majority students were seen in the CogAT 6 standardization. Those differences decreased dramatically on CogAT 7, where gaps decreased to 2-4 points for Hispanic, ELL, and FRL groups and 6 points for Black students on the CogAT 7 Alt-Verbal Battery. Differences for the Quantitative Battery were similarly improved on CogAT 7, with consistently smaller differences for all student subgroups. Notably, the gap between Asian and White students changed sign from -2 to +7.6 points. The Nonverbal Battery served as a rough control because the test had minimal changes from CogAT 6 (notably, the addition of Paper Folding). As would be expected, changes in the gaps were small with only FRL showing a smaller gap on CogAT 7.

Table 7 about here

Differences shrunk further after controlling for the effects of ethnicity, gender, FRL eligibility, ELL status, grade, and a number of school-related characteristics; particularly for ELL and FRL groups. See the estimated marginal differences in Table 7. Only the difference between Black and White students remained at a medium effect size on all three batteries.

[2] Differences observed between racial/ethnic, ELL, and FRL groups in the CogAT6 standardization data from 2000 are consistent with differences that have been observed on smaller CogAT 6 datasets in the intervening years (e.g., Lohman, Korb, & Lakin, 2008), including operational CogAT 6 datasets from 2010. Thus, it is unlikely that the differences observed between CogAT6 and CogAT7 standardization samples are due solely to changes in the U.S. population.

Discussion

This study found substantial evidence indicating that picture formats can measure general and domain-specific fluid reasoning abilities. First, the subtest intercorrelations were strong for subtests from the same battery, indicating the likelihood of domain reasoning factors in addition to Gf. This is especially convincing in the case of the Verbal Battery because the Sentence Completion format clearly draws on verbal reasoning abilities and shows strong correlations with the picture-based Verbal Analogies and Verbal Classifications subtests.

The MIRT bifactor model analyses lent further support to these findings. The model indicated the presence of a secondary domain factor for each of the three batteries, although, predictably, the influence of Gf was much stronger than any of the domain factors. We also observed, as expected, that the figural factor was quite weak. Importantly, all five picture-based formats had a substantial number of items loading strongly on the domain factor, although results were mixed for Number Analogies and Number Series with fewer items loading on the domain factor and/or showing less variance attributable to that factor relative to the general factor.

According to the MIRT results, the most impressive of the new picture formats is Picture Analogies. This format had a relatively high average domain loading, matching the Sentence Completion which we expected to be the most verbally-loaded format. Picture Classification showed very consistent but relatively weaker domain loadings. The primary drawback to the Picture Analogies and Classification formats is a lower G-loading than the more traditional Figural formats. However, it should be noted that psychometricians have spent a great deal more time refining these classic figural formats as compared to the relatively novel picture formats designed to elicit verbal or conceptual reasoning, so there is likely to be substantial room for improvement.

The quantitative test with the strongest performance was clearly Number Puzzles. At higher grade levels this test is the most school-like format and thus was expected to be the most direct test of quantitative abilities. The results for Number Series and Number Analogies were mixed, indicating that more item screening would be necessary to ensure consistent domain factor measurement. Inspection of items suggested that a subset of items in these formats can be solved using visual strategies that circumvent quantitative reasoning. Such items were good measures of Gf but often had zero or even negative loadings on the quantitative factor.

The results for the figural tests are the most difficult to interpret. The domain factor was weak and dominated by Paper Folding. The strong loading of Paper Folding items on the figural domain factor paired with the weak loadings of the Figure Matrices and Figure Classification may indicate that the domain factor for the figural formats is relatively narrow, representing either a format-specific dimension or possibly a visio-spatial factor (Gv).

As a side note, the bifactor results also highlight the value of using bifactor MIRT modeling to screen items during item development and tryout. Current test development practices tend to emphasize *g* or Gf at the expense of domain-related factors because of the use of unidimensional models and traditional item statistics during test development. Our results clearly show that when the target construct for a broad test battery involves overlapping traits, using bifactor MIRT modeling could promote greater cohesion of items with domains and help increase discriminant validity.

Our second research question addressed the convergent and discriminant validity of the picture-based CogAT 7 subtests and batteries with measures of academic achievement from the Iowa Assessments. While the first research question established evidence of a domain factor, the second set of analyses evaluated whether these domain factors represented the desired constructs

(e.g., verbal fluid reasoning that would differentially predict reading achievement outcomes). The latent regression results showed that the verbal factor is by far most valuable of the three domain factors when it comes to predicting achievement at these grade levels. This is not surprising given the oral language demands of the item formats used by the Iowa Assessments for all domains at these grade levels. As a result, regression results for the quantitative factor, which were modest as best, but are somewhat inconclusive because the Iowa Math test at these levels is heavily loaded with oral language comprehension. The validity of the figural factor is similarly weak, but results are even more inconclusive as none of the achievement measures available could be expected to draw on visual-spatial abilities (indeed most of the K-12 curriculum all but ignores these abilities, see Lohman, 2005).

Overall, subtests in the three reasoning domains do correlate strongly with academic achievement, though there is not as much differentiation in the coefficients by batteries from the same domain (verbal-reading, quantitative-mathematics) as we find for tests for older students with traditional item formats. For example, in the CogAT 7 test levels for grades 3-12, the Verbal Battery correlates r = .79 with reading achievement and r = .64 with mathematics on average, which is a larger differentiation of domain correlations than any we see at K-2 in our data. This may be a feature of the picture items or a limitation of relying on oral prompts in achievement testing for young students. Additional validity research, perhaps relying on think aloud protocols or other convergent validity evidence with ability tests, is needed to further verify that the domain factors derived from the picture-based batteries are distinctly verbal and quantitative fluid reasoning abilities and not some other construct.

Our third research question regarded the magnitude of group differences on CogAT 7. With this question, we were interested in whether the introduction of picture-based formats had

the intended effect of reducing group differences. The magnitude of group differences is important for some common uses of the CogAT such as gifted and talented identification where many school programs prefer tests that identify a gifted population that is representative of the cultural and linguistic diversity of the full student population. We found that comparisons of ELL vs. non-ELL students, FRL vs. full-price lunch, and the three ethnic/racial group contrasts all yielded small to negligible differences (below 0.4SD) between groups. These results compare favorably to the differences observed on the previous form, and are smaller than the .50-1.0 SD gaps often observed in older students (Carman & Taylor, 2010; Dickens & Flynn, 2006; Reardon & Galindo, 2009; Rushton & Jensen, 2010).

An avenue for future work may be to explore whether picture-based formats can be developed that are appropriate for older students. These items seem to be limited by increasing levels of culturally specific knowledge when test developers attempt to create more difficult items. Schulze, Beauducel, and Brocke (2005), for example, developed picture-based analogy items for their study of adult intelligence and had to rely on cultural content (e.g., recognizing Romanesque and Gothic architectural styles) that would not be considered appropriate for operational tests, which seek to avoid esoteric cultural content. Tinsley and Dawis (1972) faced similar problems in developing parallel figural and picture analogies.

Conclusion

This study found that picture-based item formats adapted from existing formats are able to measure distinct verbal, quantitative, and figural fluid reasoning abilities in young children (grades K-2). This study lends further support to the finding that Gf is best measured by sampling multiple domains to assess reasoning skills in a range of relevant contexts (Beauducel et al., 2001; Wilhelm, 2005). It also supports the use of these item formats in the assessment of

young children, for whom measuring distinctly verbal and quantitative reasoning is difficult due to their lack of literacy. Previously existing tests were limited either by excessive demands on listening comprehension, lack of differentiation of domains due to the use of a single item format, and/or the need for individual administration.

Measures of broad fluid reasoning abilities have a number of important uses in K-12 school contexts, yet the connection between intelligence theory and assessment practice can be weak. One of the key advantages of the picture-based formats is that it allows assessments to avoid over-reliance on a single test format (often figural matrices), which does not allow for capturing (or averaging out) domain effects. Beyond practical issues, another motivation for the use of this single nonverbal format with young school children is that it reduces the language load of the test. Thus, another important benefit of the picture-based item formats was the potential for reducing cultural and linguistic load of the tests while still measuring the full range of Gf skills. This study confirmed that differences between subgroups—ELL students, students on FRL, and ethnic/racial minority students—were substantially reduced while the test was able to measure distinct domain factors. Thus, picture-based formats represent a substantial improvement to assessment practice resulting from the application of intelligence theory.

References

Asparouhov, T. and Muthén B. (2010) *Bayesian Analysis Using Mplus: Technical Implementation.* Mplus Technical Report. http://www.statmodel.com

Beauducel, A., Brocke, B., & Liepmann, D. (2001). Perspectives on fluid and crystallized intelligence: Facets for verbal, numerical, and figural intelligence. *Personality and Individual Differences, 30*, 977-994.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* New York, NY: John Wiley & Sons.

Bracken, B. A. & McCallum, R. S. (1998). *Examiner's Manual: Universal Nonverbal Intelligence Test (UNIT).* Itasca, IL: Riverside Publishing

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*, 434-455.

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of the Factor-Analytic Studies*. New York: Cambridge University Press.

Cattell, R. B., & Tsujioka, B. (1964). The Importance of Factor-Trueness and Validity, Versus Homogeneity and Orthogonality, in Test Scales. *Educational and Psychological Measurement*, 24(1), 3–30.

Chen, F. F., Sousa, K. H., & West, S. G. (2005) Teacher's corner: Testing measurement invariance of second-order factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*, 471-492.

Coan, R. W. (1964). Facts, factors and artifacts: The quest for psychological meaning. *Psychological Review*, 71, 123-140.

Corno, L. (1995). The principles of adaptive teaching. In A.C. Ornstein (Ed.), *Teaching: Theory into practice* (pp. 98-115). Boston, MA: Allyn & Bacon.

Dickens, W. T., & Flynn, J. R. (2006). Black Americans Reduce the Racial IQ Gap Evidence From Standardization Samples. *Psychological Science, 17*(10), 913–920.

Corno, L., Cronbach, L. J., Kupermintz, H., Lohman, D. F., Mandinach, E. B., Porteus, A. W., et al. (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. Mahwah, NJ: Erlbaum.

DiPerna, J.C. (2005). [Review of the Cognitive Abilities Test Form 6]. In *The sixteenth mental measurements yearbook.* Retrieved from http://www.unl.edu/buros/

Dunbar, S. B., Welch, C. J., Hoover, H. D., & Frisbie, D. A. (2011a). *Iowa Assessments, Form E*. Riverside Publishing.

Dunbar, S. B., Welch, C. J., Hoover, H. D., & Frisbie, D. A. (2011b). *Iowa Assessments, Form E, Score Interpretation Guide*. Riverside Publishing.

Dunbar, S. B., Welch, C. J., Hoover, H. D., & Frisbie, D. A. (in press). *Iowa Assessments, Form E, Research and Development Guide*. Riverside Publishing.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436.

Gregory, R.J. (2004). *Psychological testing: History, principles, and applications* (4th ed.). Boston: Allyn & Bacon.

Gustafsson, J. -E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8*, 179−203.

Gustafsson, J. E. (2002). Measurement from a hierarchical point of view. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 73-95). London: Erlbaum.

Gustafsson J.E., Aberg-Bengtsson L. (2010). Unidimensionality and the interpretability of psychological instruments. In Embretson S.E. (Ed.) *Measuring psychological constructs*. Washington DC: American Psychological Association. pp. 97–121.

Gustafsson, J. E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*(4), 407–434.

Hambleton, R.K., Merenda, P.F., & Spielberger, C.D. (2005), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Holzinger, K., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*, 41–54.

Jöreskog, K. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202.

Juan-Espinosa, M., García, L.F., Colom, R. & Abad, F.J. (2000). Testing the age related differentiation hypothesis through the Wechsler's scales. *Personality and Individual Differences, 29*, 1069-1075.

Kane, H.D., & Brand, C.R. (2006). The variable importance of general intelligence (g) in the cognitive abilities of children and adolescents. *Educational Psychology, 26*(6), 751-767.

Kaufman, A. S., & Kaufman, N. L. (2004). Kaufman Brief Intelligence Test (2nd ed.). Bloomington, MN: Pearson, Inc.

Keith, T.Z., & Reynolds, M.R. (2010). Cattell-Horn-Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools, 47*(7), 635-650.

Linacre, J.M. (2006). Winsteps (Version 3.63.0) [Computer Software]. Beaverton, Oregon: Winsteps.com.

Lohman, D. F. (2000). Complex information processing and intelligence. In R. Sternberg (Ed.), *Handbook of intelligence* (pp. 285–340). Cambridge, MA: Cambridge University Press.

Lohman, D. F. (2005). The role of nonverbal ability tests in identifying academically gifted students: An aptitude perspective. *Gifted Child Quarterly, 49*, 111-138.

Lohman, D.F. (2011). *Cognitive Abilities Test, Form 7*. Rolling Meadows, IL: Riverside Publishing.

Lohman, D.F. (in press a). *Cognitive Abilities Test, Form 7, Score Interpretation Guide*. Rolling Meadows, IL: Riverside Publishing.

Lohman, D.F. (in press b). *Cognitive Abilities Test, Form 7, Research and Development Guide*. Rolling Meadows, IL: Riverside Publishing.

Lohman, D. F., & Hagen, E. (2001). *Cognitive Abilities Test, Form 6*. Itasca, IL: Riverside.

Lohman, D. F., & Hagen, E. (2002). *Cognitive Abilities Test, Form 6: Research handbook*. Itasca, IL: Riverside.

Lohman, D. F., Korb, K., & Lakin, J. (2008). Identifying academically gifted English language learners using nonverbal tests: A comparison of the Raven, NNAT, and CogAT. G*ifted Child Quarterly*, *52*, 275-296.

Lohman, D. F., Gambrell, J. (2012).Using Nonverbal Tests to Help Identify Academically Talented Children. *Journal of Psychoeducational Assessment*, *30*, 25-44.

Messick, S. (1989). Validity. In R. L. Linn (Ed*.). Educational measurement* (3rd ed.; pp. 13-103). New York: Macmillan.

Muthén, L. K., & Muthén, B. (2010). *Mplus: Statistical analysis with latent variables*. Los Angeles: Muthén & Muthén.

Reardon, S. F., & Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *American Educational Research Journal, 46*(3), 853–891.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor Models and Rotations: Exploring the Extent to which Multidimensional Data Yield Univocal Scale Scores. *Journal of Personality Assessment, 92*(6), 544–559. doi:10.1080/00223891.2010.496477

Rijmen, F. (2009). *Three multidimensional models for testlet-based tests: Formal relations and an empirical comparison* [ETS RR-09-37]. Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12*, 1151-1172.

Rushton, J. P., & Jensen, A. R. (2010). The rise and fall of the Flynn Effect as a reason to expect a narrowing of the Black-White IQ gap. *Intelligence, 38*(2), 213–219.

Schneider, W.J., & McGrew, K.S. (2012). The Cattell-Horn-Carroll model of intelligence. In D.P. Flanagan & P.L. Harrison (Eds.), Cont*emporary Intellectual Assessment: Theories, Tests, and Issues* (3rd ed.; pp. 99-144). New York, NY: Guilford Press.

Schulze, D., Beauducel, A., Brocke, B. (2005). Semantically meaningful and abstract figural reasoning in the context of fluid and crystallized intelligence. *Intelligence, 33*, 143-159.

Snow, R.E., & Lohman, D.F. (1984). Toward a theory of cognitive aptitude for learning from instruction. *Journal of Educational Psychology, 76*(3), 347-376.

Starkey, P. (1992). The early development of numerical reasoning. *Cognition, 43*, 93-126.

Süß, H., & Beauducel, A. (2005). Faceted models of intelligence. In O. Wilhelm, & R. W. Engle (Eds.), *Handbook of Understanding and Measuring Intelligence* (pp. 313-332). Thousand Oaks, CA: Sage.

Thomson, G. (1948). *The factorial analysis of human ability*. New York: Houghton Mifflin Co.

Tinsley, H.E.A., & Dawis, R.V. (1972). *The Equivalence of Semantic and Figural Test Presentation of the Same Items* [Monograph]. (ED068515). Retrieved from http://www.eric.ed.gov/

Wilhelm, O. (2005). Measuring reasoning ability. In O. Wilhelm & R. W. Engle, *Handbook of understanding and measuring intelligence* (pp. 373-392). London: Sage Publications.

Yung, Y., Thissen, D., & McLeod, L. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64*, 113–128.

Table 1

Characteristics of the Sample (Percentages)

	CogAT Only	CogAT + ITBS
Number of Students	18,153	6,320
Female	48.6	48.2
Grade		
K	32.3	-
1	30.7	50.0
2	37.0	50.0
Region		
Northeast	8.8	13.9
Midwest	12.9	11.6
South	51.5	32.8
West	26.8	41.6
Ethnicity		
White	55.6	54.6
Asian	4.5	5.2
Hispanic	18.3	17.3
Black	15.6	14.5
Native American	1.9	3.0
Pacific Islander	1.0	0.8
English-Language Learner	4.2	2.8
Asian	0.7	24.4
Hispanic	3.0	57.4
Other	0.5	18.2
Free-Reduced Lunch	21.4	27.4
White	7.7	10.6
Asian	0.4	0.8
Hispanic	7.9	10.2
Black	4.3	3.9

Table 2

Raw score descriptive statistics and alpha reliabilities

Grade	K				1				2			
CogAT Level	5/6				7				8			
IA Level	-				6				7			
	Mean	*SD*	*# of Items*	*α*	*Mean*	*SD*	*# of Items*	*α*	*Mean*	*SD*	*# of Items*	*α*
Picture Analogies	8.49	2.21	14	.55	10.59	2.53	16	.65	10.59	2.53	18	.72
Picture Classification	7.46	2.49	14	.58	9.95	2.5	16	.61	9.95	2.5	18	.64
Sentence Completion	9.29	2.28	14	.56	10.61	2.41	16	.58	10.61	2.41	18	.61
Number Analogies	5.77	2.79	14	.69	8.12	2.98	16	.74	8.12	2.98	18	.75
Number Series	5.94	2.12	14	.44	7.42	2.54	16	.58	7.42	2.54	18	.74
Number Puzzles	4.42	2.21	10	.63	6.14	2.88	12	.78	6.14	2.88	-	-
Figure Matrices	7.53	2.9	14	.72	9.58	2.64	16	.67	9.58	2.64	18	.67
Figure Classification	8.01	3.04	14	.74	9.90	2.65	16	.65	9.90	2.65	18	.66
Paper Folding	5.46	2.03	10	.53	7.48	2.14	12	.67	7.48	2.14	16	.8
Verbal	25.25	5.32	42	.75	31.15	5.87	48	.8	31.15	5.87	54	.83
Alt-Verbal	15.95	3.88	28	.67	20.54	4.29	32	.75	20.54	4.29	36	.79
Quantitative	16.13	5.49	38	.77	21.68	6.73	44	.85	21.68	6.73	50	.88
Figural	21.01	6.52	38	.84	26.96	6.02	44	.82	26.96	6.02	52	.87
Q-N Composite	40.27	11.90	76	.88	48.64	11.71	88	.90	57.50	14.22	102	.92
IA Vocabulary	-	-	-	-	18.09	3.49	27	.63	17.97	5.59	26	.89
IA Reading Comp	-	-	-	-	21.09	7.97	34	.9	26.86	6.72	35	.91
IA Language	-	-	-	-	21.44	4.76	31	.57	21.94	6.00	34	.88
IA Word Analysis	-	-	-	-	28.07	4.27	33	.84	25.45	4.64	32	.81
IA Listening	-	-	-	-	17.71	4.12	27	.72	19.03	4.23	27	.77
IA Mathematics	-	-	-	-	24.51	5.09	35	.84	27.86	5.74	41	.84
IA Math Computation	-	-	-	-	-	-	-	-	18.85	4.63	25	.84
IA Reading Composite	-	-	-	-	39.25	9.80	61	.89	44.96	11.66	61	.94
IA Math Composite	-	-	-	-	-	-	-	-	46.80	9.33	66	.90

Table 3

Correlations between CogAT subtests across grades K-2

	PA	PC	SC	NA	NS	NP	FM	FC	PF
Picture Analogies	(.73)	.61	.55	.65[a]	.50	.51	.55[a]	.53	.56
Picture Classification	.87	(.67)	.55	.60	.47	.52	.54	.58[a]	.55
Sentence Completion	.81	.85	(.62)	.54	.44	.49	.47	.50	.50
Number Analogies	.85[a]	.82	.78	(.79)	.55	.57	.62[a]	.57	.59
Number Series	.70	.70	.67	.75	(.68)	.51	.53	.51	.52
Number Puzzles	.67	.72	.70	.73	.71	(.52)	.52	.50	.52
Figure Matrices	.76[a]	.79	.72	.84[a]	.78	.70	(.70)	.63	.57
Figure Classification	.74	.86[a]	.77	.78	.75	.68	.91	(.69)	.56
Paper Folding	.75	.77	.73	.77	.72	.67	.78	.77	(.76)

*Note*s. Cross-grade reliabilities on the diagonal. The upper diagonal shows observed correlations, lower diagonal shows correlations corrected for unreliability. [a] Correlations between tests that share format (analogies or classification).

Table 4

Bifactor MIRT results

	Percentage of items loading on domain factor [a]	Average Item Variance Components [b]					
		All Items			Items with Significant Domain Loadings		
		Gf	Domain	Domain % [c]	Gf	Domain	Domain % [c]
Picture Analogies	88	19	6	23	21	7	26
Picture Classification	94	18	3	15	18	3	15
Sentence Completion	88	13	4	24	15	5	25
Number Analogies	56	23	0	1	11	5	30
Number Series	69	16	3	16	16	6	28
Number Puzzles	100	32	9	22	32	9	22
Figure Matrices	25	29	1	2	34	13	28
Figure Classification	41	27	0	1	38	3	7
Paper Folding	81	28	6	18	30	9	23

[a] The percent of items showing a statistically significant positive domain factor loading at $p < .001$

[b] Computed by squaring the average standardized loading

[c] This is the percent of reliable variance accounted for by the domain factor. Equal to D/(Gf+D)

Table 5

Standardized Regression coefficients between Latent Factors and Iowa scales

Factor	Vocabulary	Reading	Listening	Math
General	.594*	.661*	.639*	.747*
Verbal	.360*	.256*	.384*	.310*
Quantitative	.098*	.088*	.084*	.138*
Figural	.024	.029	.066*	.050*

* $p < .001$

Table 6

Disattenuated Correlations between CogAT-7 Item Formats and ITBS-E Scales, Grades K-2

Item Format	Vocabulary	Reading	Listening	Math	Math Computation[a]
Picture Analogies	.64	.62	.66	.71	.48
Picture Classification	.65	.60	.67	.70	.51
Sentence Completion	.76	.65	.84	.83	.60
Number Analogies	.62	.65	.68	.77	.57
Number Series	.51	.49	.55	.65	.60
Number Puzzles[b]	.52	.48	.56	.64	N/A
Figure Matrices	.53	.52	.58	.66	.58
Figure Classification	.54	.52	.61	.65	.55
Paper Folding	.58	.57	.65	.70	.52
Battery scores					
Verbal USS	.79	.73	.82	.86	.60
Alt-Verbal USS[c]	.73	.69	.75	.79	.55
Quantitative USS	.69	.69	.73	.83	.65
Figural USS	.65	.65	.70	.77	.59
Q-N Composite USS	.69	.69	.74	.83	.65

Note: All correlations are significant at $p < .001$. N = approx. 6,000 unless otherwise noted. [a]Not administered at grades K and 1, N = approx. 3,000. [b]Grade 2 data not included because the format changes from pictures to numbers, N = approx. 3,000. [c]Alt-Verbal does not include the Sentence Completion subtest which requires receptive language.

Table 7

Mean Differences Between Focal Groups - SAS Points[a]

	ELL	FRL	Hispanic	Black	Asian
C7 Sample Sizes	763	3,878	3,323	2,826	821
C6 Sample Sizes	3,190	15,141	5,000	6,651	1,192
C7 Alt-Verbal	-2.3	-4.2	-2.2	-5.9	6.0
C6 Verbal	-11.4	-11.1	-15.7	-14.3	-7.6
C7 Quantitative	-2.8	-5.0	-4.7	-7.2	7.6
C6 Quantitative	-8.9	-10.0	-12.7	-13.0	-2.0
C7 Nonverbal	-4.6	-5.0	-4.3	-7.1	5.8
C6 Nonverbal	-4.3	-8.4	-6.8	-10.0	5.7
Estimated Marginal Differences[b]					
C7 Alt-Verbal	*-1.5*	*-2.9*	*-1.9*	*-6.6*	*4.2*
C7 Quantitative	*-0.1*	*-3.8*	*-4.5*	*-7.6*	*5.4*
C7 Nonverbal	*-2.4*	*-3.5*	*-3.0*	*-7.5*	*4.7*

[a] SAS scale is $M = 100$, $SD = 16$. [b] Marginal mean differences were estimated his analysis was carried out using a general linear model controlling for the main effects of ethnicity, gender, FRL, ELL, age, grade, school SES, school size, region of the country, and public vs. private school.

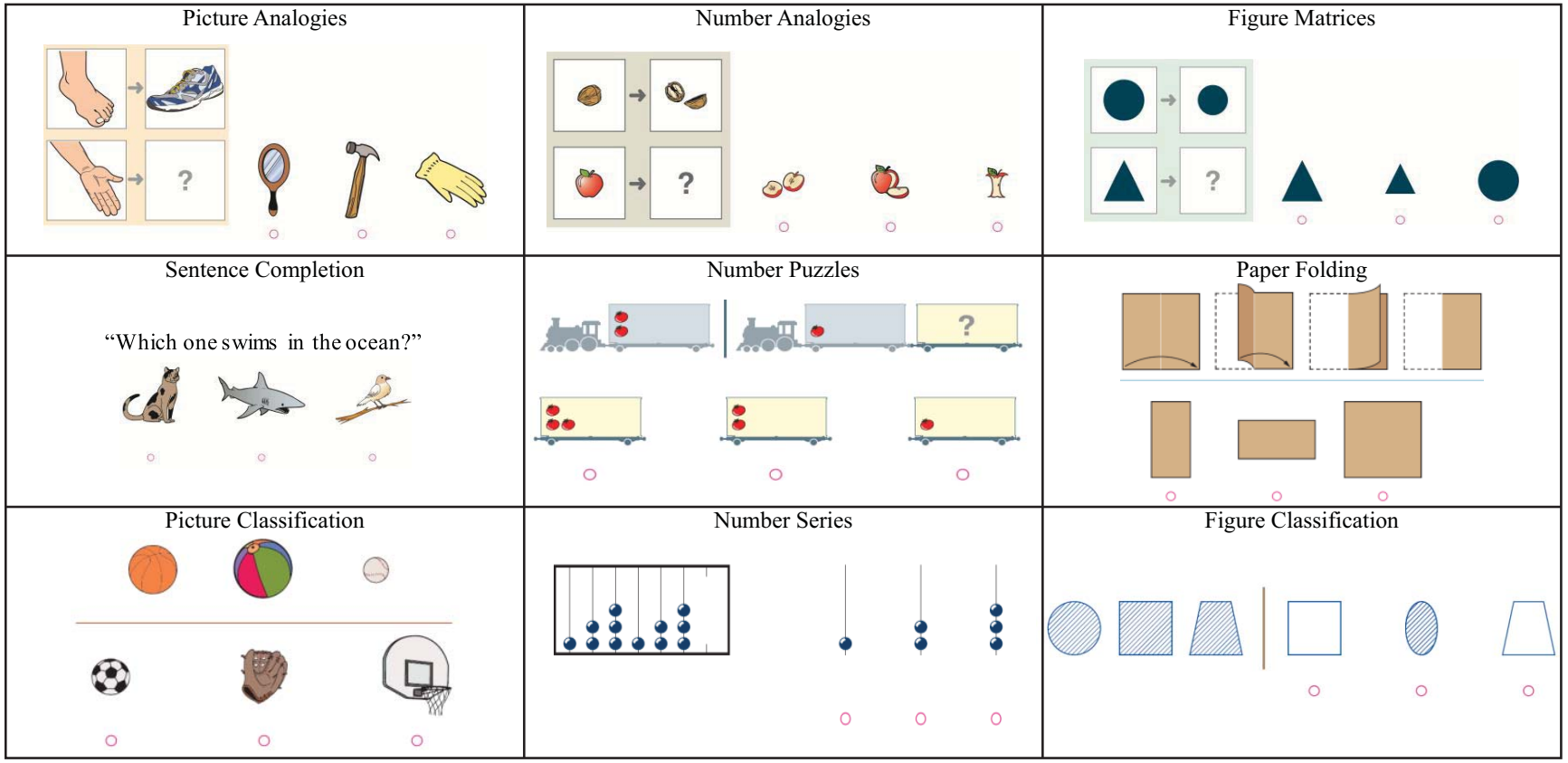

Figure 1. CogAT 7 Item Formats. Picture Reasoning subtests on Form 7 of the Cognitive Abilities Test (Lohman, 2011) showing examples of item formats for grades K–2.

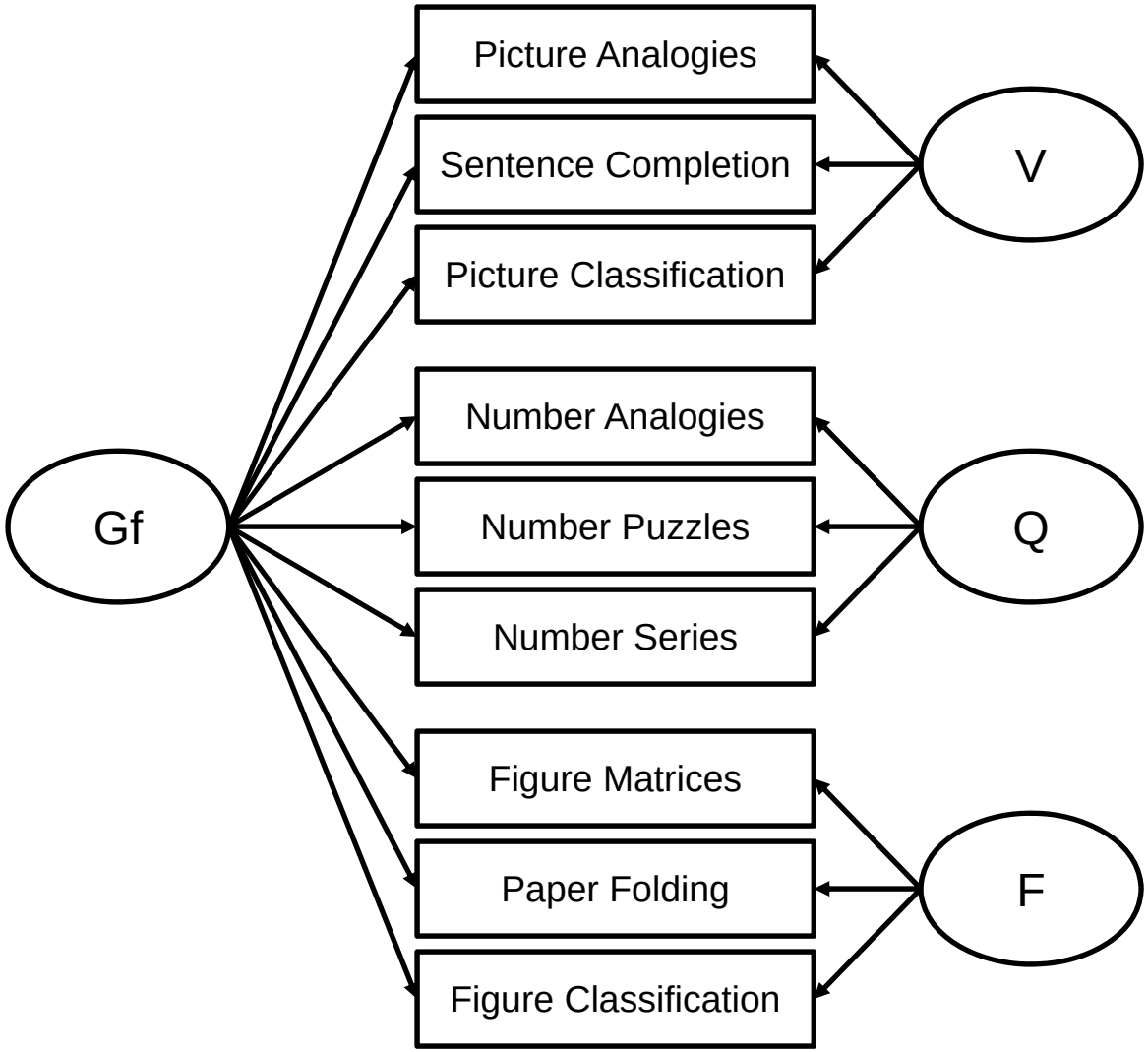

Figure 2. Bifactor 2PL MIRT model. Note that subtest blocks represent items rather than raw scores.