Psychometric Evaluation of Aurora-*a*: An Augmented Assessment of Analytical, Practical, and Creative Abilities in Middle Childhood and Early Adolescence

Sascha Hein

Mei Tan

Yusra Ahmed

Julian G. Elliot

David Bolden

Robert J. Sternberg

Elena L. Grigorenko

Sascha Hein, Department of Education and Psychology, Freie Universität Berlin, and Child Study Center, Yale University; Mei Tan, HIV Center for Clinical and Behavioral Studies, Columbia University; Yusra Ahmed, Texas Institute of Measurement, Evaluation, and Statistics, University of Houston; Julian G. Elliot, School of Education, Durham University; David Bolden, School of Education, Durham University; Robert J. Sternberg, College of Human Ecology, Cornell University; Elena L. Grigorenko, Department of Psychology, University of Houston, and Child Study Center, Yale University, and Baylor College of Medicine.

Address correspondence to Elena L. Grigorenko, Department of Psychology, University of Houston, 4349 Martin Luther King Boulevard, Houston, TX 77204-6022, Phone: 1-713-743-0595, Email: Elena.Grigorenko@times.uh.edu.

Authors' email addresses: Sascha Hein, Sascha.Hein@fu-berlin.de; Mei Tan, mtt2114@cumc.columbia.edu; Yusra Ahmed, Yusra.Ahmed@times.uh.edu; Julian G. Elliot, joe.elliott@durham.ac.uk; David Bolden, d.s.bolden@durham.ac.uk; Robert J. Sternberg, rjs487@cornell.edu, Elena L. Grigorenko, Elena.Grigorenko@times.uh.edu

Abstract

The Theory of Successful Intelligence defines intelligence as the integrated set of abilities and competencies in specific domains needed to attain success in life. Informed by this theory, we examined the dimensionality, reliability, and validity of an augmented intelligence test, Aurora-*a*, a 17-subtest assessment of analytical, practical, and creative abilities and figural, numerical, and verbal competencies in middle childhood and early adolescence. Using data from 3470 students (1808, or 52.1%, identified as male) from the United Kingdom and the United States, we found support for the unidimensionality and adequate reliability of the 17 subtests. An exploratory structural equation model outperformed confirmatory factor analysis on goodness-of-fit, theory alignment, model parsimony, and interpretability, illustrating the multifaceted nature of items assessing analytical, practical, and creative abilities. Weak to strong correlations (ranging r = .20 to .72) with criterion assessments of academic performance corroborated the validity of Aurora-*a*.

Keywords: Cognitive assessment, analytical thinking, creativity, practical ability, ESEM

Psychometric Evaluation of Aurora-*a*: An Augmented Assessment of Analytical, Practical, and Creative Abilities in Middle Childhood and Early Adolescence

Contemporary research explicitly states that intelligence cannot be adequately captured as a single high-order general factor (e.g., Spearman's theory of general intelligence, or g; the hierarchical Cattell-Horn-Carroll theory of cognitive abilities; Carroll, 1993; Spearman, 1904), as it encompasses a range of specific cognitive processes, abilities, and skills that interact dynamically across different developmental stages and sociocultural contexts (Breit et al., 2021; Demetriou et al., 2022; Gardner, 2011; Sternberg, 2012, 2020). Relevant thinking has unfolded in multiple ways, focusing both on the developmental transformation of abilities that lead to the ontogenetic maturation of intelligence and the phylogenetic typology of cognitive processes that substantiate human abilities.

Regarding the former, there have been several theoretical attempts to integrate psychometric *g* and developmental frameworks of intelligence to understand the extent to which individual differences in performance on cognitive tests differentiate across the lifespan. Thus, the integrated differential-developmental theory of (general) intelligence (Demetriou et al., 2018; Demetriou et al., 2022; Demetriou et al., 2023) suggests that the cognitive profile of *g* changes significantly as different cognitive processes become more prominent at various developmental stages. This theory recognizes the widely accepted construct of *g* in psychometrics, where *g* represents the positive correlations between various cognitive tasks as outlined in hierarchical models such as the CHC model. The structural equation modeling (SEM) results presented by Demetriou et al. (2022) confirm the importance of distinct cognitive factors like attention control, working memory, and reasoning in overall cognitive functioning during the developmental period of ages 10-12 years. During this period, these cognitive processes become more integrated with *g*, reflecting their increasing complexity and importance in cognitive functioning. In contrast, attention control, while still relevant, shows less variation and impact on *g* compared to earlier developmental stages.

The differentiation hypotheses of intelligence (Breit et al., 2022; Breit et al., 2021; Breit et al., 2024) also suggest that relationships between *g* and specific cognitive abilities change with age, and these changes vary across different levels of ability. However, the evidence is not consistent across all abilities and age groups. For example, For example, findings suggest that numerical reasoning abilities exhibit the most pronounced differentiation as children age, followed by verbal reasoning abilities, while figural reasoning abilities remain relatively stable in their correlations with other cognitive abilities (Breit et al., 2021). Further, high-ability children tend to show greater differentiation of specific cognitive skills at earlier ages, suggesting that their cognitive development follows a more accelerated and specialized trajectory. This contrasts with lower-ability children, who may exhibit a more gradual and integrated development of cognitive skills. This suggests that the developmental trajectories of cognitive skills are complex and may not fit neatly into a single model of differentiation. It also suggests that the landscape of correlations between different cognitive abilities might impact the *g* factor differentially at different ability levels and at different developmental stages.

These frameworks offer pertinent developmental considerations reflecting the lifespan dynamics of intellectual development but do not exemplify in detail the content of the developmentally transforming abilities, that is, leaving room for various typologies of such abilities. The Theory of Successful Intelligence, TSI, exemplifies a theoretical approach focusing on such typology and advances a broad developmental definition of cognitive abilities, competencies, and expertise, supported by evidence indicating that intelligence is a multifaceted construct going beyond conventional conceptions of the *g*-factor (Sternberg, 2020). Notably, the TSI is one of several theories that expand traditional conceptions of intelligence by highlighting other abilities that have historically been understudied and underappreciated. Another theory based on a similar notion is the theory of multiple intelligences proposed by Howard Gardner (e.g., Gardner, 2006). There are other, more recent attempts to better understand the variety and scope of human (cognitive) abilities (e.g., mutualism theory; van der Maas et al., 2006) and their expression in different (e.g., harsh and unpredictable) contexts and environments (Ellis et al., 2022).

Successful intelligence is defined as the integrated set of abilities, competencies, and expertise needed to attain success in life. Successfully intelligent people adapt to, shape, and select environments through a balance in their use of analytical, creative, and practical abilities. Hence, analytical, creative, and practical abilities are equally important to intellectual functioning (Sternberg, 2020). Analytical intelligence involves analyzing, evaluating, judging, and comparing and contrasting. Analytic abilities are exhibited in reasoning and logical thinking as exercised in activities such as persuasive writing, debating, analyzing, and mathematical problem-solving. These abilities are typically evaluated, in part, by conventional tests of intelligence. Practical intelligence is involved when individuals apply their abilities and knowledge to the problems that confront them in daily life, for example, at work or home. Practical abilities are exercised in leadership and other social interactions and in the adaptation and application of knowledge in real-world problem-solving (Sternberg & Grigorenko, 2000). Creative abilities are reflected in the capacity to generate new ideas, create, imagine, invent, and design in activities like writing, drawing, building, and imaginative play. Creative intelligence is particularly well captured by tasks and situations assessing how well an individual copes with relative novelty and cognitive ambiguity. The TSI has been studied from a developmental perspective, tracking the transformation of abilities into competencies and expertise (Sternberg & Grigorenko, 2012). Its definition of intelligence as overlapping sets of cognitive skills composed to execute problem-solving in different ways (i.e., analytically, creatively, and/or practically) implies that the theory may be extended across the lifespan. This has been illustrated in various studies focusing on school-age children (Grigorenko et al., 2002; Grigorenko et al., 2004; Sternberg et al., 2001), college and professional students (Hedlund et al., 2006), and adults (Grigorenko & Sternberg, 2001; Sternberg et al., 2000).

To illustrate, Sternberg and the Rainbow Project Collaborators (2006) utilized a diverse set of assessments designed to quantify analytical, creative, and practical abilities to examine their associations with college GPA and verbal and math SAT scores. The authors found that some of these assessments (e.g., a latent variable capturing general practical abilities) were significantly related to college GPA over and above high-school GPA and SAT scores. In contrast, other assessments showed no such incremental prediction (e.g., the performance-based indicators of creativity). The psychometric and confirmatory factor analyses supported single-factor models for creative and practical abilities. The Rainbow Project was conducted with "students predominantly in their first year of college or their final year of high school" (Sternberg & The Rainbow Project Collaborators, 2006, p. 326).

Stemler and colleagues (Stemler et al., 2006; Stemler et al., 2009) incorporated an assessment of analytical, creative, practical, and memory skills (i.e., different facets of successful intelligence expressed in specific items gathered in specific subscales) in the context of Advanced Placement (AP) Psychology and Statistics tests given to seniors in high school. Students' scores on the TSI subscales in the AP tests correlated as expected with actual AP tests, with the memory TSI subscale correlating the most and creative skills correlating the least highly with the actual AP test score. In addition, Q-factor analysis revealed distinct profiles of strengths and weaknesses across the four (analytic, creative, practical, and memory) facets of successful intelligence in both AP Psychology and AP Statistics test performance; some of these strengths would otherwise not have been revealed in the original AP test items. The TSI-augmented AP exams also reduced differences in ethnic group mean scores on both the AP Psychology and Statistics exam, a key goal of the test augmentation.

At the other end of the age spectrum, the Bilingual Early Language-Learner Assessment (BELLA) was developed to assess pre-kindergarteners early literacy, numeracy, science, and social-emotional simultaneously with their analytical, creative, and practical/social skills in over 700 test items (Tan et al., 2023). Data collected on BELLA items with 3-6-year-old children showed that when analytical, creative, and practical items were clustered across knowledge domains, pass rates (numbers of items correct) indicated a consistent increase in each TSI skill by age, particularly on the creative items (Tan et al., 2023).

The Aurora Battery operationalized the TSI to address issues of diversity in the identification of gifted children in the United States, a process typically carried out in the 4th to 6th grades (Chart et al., 2008; Speirs Neumeister et al., 2007; Sternberg & Grigorenko, 2002), thus adding to the range of developmental stages the TSI has empirically addressed. It has often achieved this by incorporating cognitive ability assessment into various age- or grade-determined levels of content knowledge. Aurora is the first TSI assessment to focus on cognitive ability alone, to encourage diversity in gifted education at the primary school level.

In the current study, the focus on a truncated age range (10-12-year-olds) constrains our ability to fully explore the developmental changes and differentiation of cognitive abilities over a broader span of childhood and adolescence. Consequently, the primary aim of this study is to evaluate the psychometric properties of the Aurora Battery at a single point in time, rather than through dynamic and longitudinal analyses. Prior research indicates that the nature of intelligence evolves significantly across the developmental span, underscoring the necessity of both cross-sectional and longitudinal investigations to comprehensively understand the progression and transformation of cognitive abilities over time. Constructs of the TSI have not yet been studied longitudinally. Yet, the theory has demonstrated its relevance and utility in multiple cross-sectional studies covering a substantial chunk of lifespan.

The Aurora Battery of Intelligence

As noted above, the Aurora Battery was originally conceived as a supplement to current methods of gifted identification generally applied in the United States (US), which traditionally have consisted mainly of IQ-based assessments that prioritize analytical skills and memory (e.g., Hodges et al., 2018). Aurora was developed based on the TSI and was initially designed to inclusively identify intellectually gifted children from about 9 to 12 years of age in the US, correcting for diversity biases (e.g., Ricciardi et al., 2020) in the identification based on conventional intelligence tests. Aurora deliberately targets analytical, practical, and creative abilities across figural, numerical, and verbal domains of students in grades 4 through 6. It also presents a range of response types, from multiple choice to open-ended short and long answers. By assessing various intellectual strengths, Aurora evaluates patterns of performance across a broader and unconventional range of skills assessed by traditional instruments (Chart et al., 2008; Tan et al., 2009). Supplemental Table 1 contains additional information about each of the Aurora-*a* subtests. The psychometric properties of the English version of Aurora-*a* have not previously been systematically examined. The present study sought to fill this gap.

Previous Research on Aurora-*a*

Aurora-*a* has been translated into several languages and adapted in several countries around the globe. Published research has predominantly focused on analytic approaches to confirm the factor structure of Aurora-*a* (e.g., CFA). Some studies aimed to determine the validity of the Aurora-*a* subtest scores, for example, by interrogating correlations with measures of academic achievement. Here, we briefly summarize past research on Aurora-*a* before presenting the aims of the present study.

Evidence of the Internal Factor Structure. A 2016 study assessed 400 randomly selected children (50% girls) aged 9-12 years from 24 schools in Isfahan, Iran, using an adapted version of Aurora-*a* (Aghababaei et al., 2016). The authors used 16 of the 17 Aurora-*a* subtests; the *Shapes* subtest was not used. They used subtest total scores to estimate a second-order CFA and found that this model fit their data well after freely estimating the covariance between two error terms ($\chi^2 = 28.23$, $p = .001$; RMSEA = 0.06; CFI = 0.94; TLI = 0.93). The composite scores of analytical, practical, and creative abilities were suitably reliable, with Cronbach's α of 0.83, 0.70, and 0.88, respectively. However, the unidimensionality and reliability of each subtest and inter-rated reliability for open-ended items were not reported. Similarly, factor loadings were not reported. Finally, alternative models were not evaluated.

A recent study from Turkey (Aslan & Soysal, 2021) administered Aurora-*a* to 520 randomly selected students aged 9-12 years who attended public and private schools in Istanbul. Three of the 17 subtests (i.e., *Silly Headlines*, *Decisions*, and *Figurative Language*) were excluded from the factor model as they were found to be poorly understood by the study participants. The authors applied a CFA model to a dataset that included 115 observed indicators (i.e., subtest items) after eliminating one item from the *Paper Cutting* subtest and three items from the *Shapes* subtest due to standardized pathway coefficients below 0.30. Unidimensionality was appraised for each subtest. The analytical subtests met the conventional cut-offs of several model fit indices (i.e., CFI ≥ 0.95) except for the RMSEA estimates that were relatively high (0.08 and above) for four out of six subtests (*Metaphors, Homophone Blanks, Letter Math,* and *Floating Boats*). For practical ability, three of the four applied subtests showed adequate fit to the data. The *Maps* subtest did not fit the data well. Finally, for creative ability, the parameter estimates showed high RMSEA values ranging from 0.15 to 0.23, indicating a poor fit to the data. The reliability coefficients ranged from 0.64 to 0.92, with the lowest estimate found for the *Shapes* subtest (Cronbach's α = 0.64). A correlated three-factor model fit the data well.

In a study with students enrolled in summer enrichment programs in Saudi Arabia, Ayoub and Aljughaiman (2016) reported the results of a CFA to appraise the factor structure of Aurora-*a*, which had been administered to a sample of 442 students. The authors found a model with three correlated factors reflecting analytical (Cronbach's α = 0.88), practical (Cronbach's α = 0.85), and creative abilities (Cronbach's α = 0.82) to fit their data reasonably well (χ^2/df = 1.12; RMSEA = 0.036; GFI = 0.97; AGFI = 0.95, NFI = 0.93). Details on the model specifications and parameter estimates were not reported. In an earlier study (Aljughaiman & Ayoub, 2012) with fewer students (n = 196), the same authors reported the results of a CFA that showed an adequate fit of the three-factor model (χ^2 = 34.99, p = .069; RMSEA = 0.048; GFI = 0.96; AGFI = 0.93, NFI = 0.97). However, the authors computed nine composite scores (e.g., analytical-numerical) for the CFA instead of using the subtest scores or item-level data, thereby limiting the comparability with the published research on Aurora-*a*.

In a study conducted in the Netherlands, Aurora-*a* was administered to 499 children from fourth to sixth grade (Gubbels et al., 2016). The authors did not use the *Silly Headlines* subtest because they deemed it “problematic to maintain equivalencies with respect to meaning, psychometric construct, and item difficulty” (Gubbels et al., 2016, p. 228). Inter-rater reliability for open-ended items was reported to range between $r = 0.72$ and 0.95, and agreement ranged between 66% (Book Covers) and 77.4% (Multiple Uses). A total of 9 items were excluded from the factor and reliability analyses for various reasons (e.g., low item-total correlations). Low reliability (greatest lower bound, GLB) prompted the authors to exclude the *Shapes* subtest (measuring analytical ability; GLB = 0.39) from the analyses. The 15 remaining subtests were then subjected to a CFA. A correlated three-factor model showed a poor fit to the data with low CFI (0.88) and high RMSEA (0.09) estimates. In this model, analytical and practical abilities correlated at $r = 1.00$. Based on this finding, the authors combined analytical and practical abilities into one factor and tested the fit of this “adapted two-factor model,” which showed the best fit compared to alternative models ($\chi^2 = 325.81$, $p < .001$; RMSEA = 0.08; CFI = 0.92).

Aurora-*a* has also been translated and adapted for use with 431 8-15-year-olds in the Murcia region in Spain (Prieto et al., 2015). The authors analyzed correlations between Aurora-*a* ability and domain scores with a total score across all subtests (ranging from .717 to .877). Moreover, the Aurora-*a* scores were correlated with non-verbal general intelligence, as assessed by the Spanish version of the Cattell test of *g*, scale 2. Results showed the strongest correlation between *g* and practical ability ($r = .651$, $p < .001$) and the weakest correlation between *g* and creative ability ($r = .503$, $p < .001$). Other psychometric properties were not reported.

Finally, Tan and colleagues (2012) described and justified the perception of shadow orientation as a figural componential facet of practical intelligence. This component is assessed in the *Toy Shadows* subtest, in which children are shown views of a toy from different angles and then shown a light shining on the toy. Given the toy’s orientation in the light, children must select the shadow cast by the toy. CFA with data from Greece, Saudi Arabia, and the US showed configural invariance of the subtest across the three countries (RMSEA = .013) and measurement equivalence indicated by corresponding factor loadings across all country models (ΔCFI = 0). That is, the subtest's properties remained consistent across countries. In addition, although the validity of this subtest has yet to be explored, low correlations of *Toy Shadows* performance with analytical (*r* = .26 to .28) and creative (*r* = -0.03) subtests of Aurora indicate that the subtest may capture a distinct set of practical abilities.

Evidence of Validity. There have been several attempts to gauge the convergent and divergent validity of Aurora-*a*. For instance, Mandelman and colleagues (2013) collected Aurora-*a* data from 145 4th to 6th-grade students in a suburban Midwestern US private, parochial school. The authors also administered the TerraNova (CTB/McGraw-Hill, 2010) to collect data on students' performance in reading, language, and mathematics tests. Partial correlations (controlling for administration delay between both measures) between Aurora-*a* scores and the TerraNova scores ranged from 0.30 (*p* < .001) for the correlation between creative ability and the TerraNova language scores to 0.71 (*p* < .001) for the correlation between practical ability and math performance. In a subsequent study with data from the same sample (Mandelman et al., 2016), the authors demonstrated that Aurora-*a* scores predicted baseline (December grading period) GPA scores. However, the predictive effect was more limited regarding overall GPA growth from December to June. Compared with overall GPA and various subjects, the predictive power of Aurora-*a* was strongest for the increase in science grades, particularly concerning practical ability (β = .50) but also creative (β = .22) and analytical ability (β = .12). The statistical significance (*p*-values) of these coefficients was not reported.

Another study (Mourgues, Tan, et al., 2016) used Aurora's five creativity subtests in a sample of 1165 7th graders from the UK to predict future academic performance as well as the mediating role of creativity in the association between two academic performance tests taken over more than five years. Two measurement models showed adequate fit to the data: a correlated factors model with two additional latent factors reflecting creativity in the verbal and figural domains (χ^2 = 812.26, p < .001; RMSEA = 0.014; CFI = 0.984) and a second-order factor model for creativity and five factors, one for each creativity subtest (χ^2 = 990.11, p < .001; RMSEA = 0.019; CFI = 0.97). Structural model analysis showed that the total creativity factor was moderately related to writing, math, and science scores as indexed by the Key Stage-2 tests, that is, standardized academic achievement tests used at the end of primary education (age 11 years) in the UK. Aurora's creativity scores also predicted students' performance on the General Certificate of Secondary Education (GCSE), a national examination taken by secondary students (usually at the age of 16 years) in eight or nine subjects (β estimates ranged from .16, p < .05, for GSCE Science scores to .25, p < .01, for GCSE English scores).

Finally, several studies focused on gifted identification with Aurora-*a* are beyond the scope of the present study and will not be described in detail here. Interested readers may refer, for example, to studies by Ferrando and colleagues (2016), who conducted a study in the Murcia region of Spain to examine profiles of gifted and talented students; Mandelman and colleagues (2013), who addressed gifted identification in the US using Aurora-*a* and the TerraNova assessment; Kornilov and colleagues (2011), who used data from the UK to identify gifted children; and Tan and colleagues (2013), who examined the performance of Aurora's *Metaphors* subtest to evaluate its ability to identify gifted students. One conclusion from these studies is that Aurora-*a* complements other methods of gifted identification.

Study Aims

As reviewed above, several studies have demonstrated the psychometric properties of Aurora-*a*. Despite the strengths of previous studies, several issues and open questions still need to be addressed. For instance, previous research that has examined the factor structure of Aurora-*a* (e.g., studies from Iran, Netherlands, and Turkey) has yet to report evidence of (criterion) validity. The studies that have examined evidence of validity (e.g., Mandelman et al., 2013, using data from the US) relied on small samples that precluded a thorough psychometric analysis. Unidimensionality of the 17 Aurora-*a* subtests is another aspect that has not been appraised in several previous studies (e.g., studies from Iran, Saudi Arabia, and Spain). Finally, inter-rater reliability for open-ended items was not reported in most reviewed research studies (e.g., studies from Iran, Spain, and Turkey).

Based on Aurora-*a*'s theoretical underpinning, this study sought to address, at least in part, the described limitations of previous research on Aurora-*a* and to empirically determine the factor structure of the 17 Aurora-*a* subtests in a sample of fourth to sixth-graders in the United Kingdom (UK) and the US. The first aim was to test the unidimensionality and reliability of each Aurora-*a* subtest using CFA. The second aim was to utilize the 17-subtest composite scores to test the factor structure of Aurora-*a* as a whole (i.e., including abilities and domains in the same model). Similar analytic approaches have been presented in the literature using translated and adapted versions of Aurora-*a*. Here, we estimated the fit of nine models to the data, the first six of which were tested using CFA. Specifically, the first model comprised a general factor accounting for variability in the 17 subtests. The second model specified three latent factors representing analytical (six subtests), practical (six subtests), and creative ability (five subtests). We hypothesized that this model would better fit the data compared with the general factor model. The third model contained three latent variables to capture domain-specific variations in the subtests based on the figural (six subtests), numerical (five subtests), and verbal (six subtests) representation of the items. We hypothesized that the third model would yield a similar fit to the data as the second model and a better fit than the general-factor model. Re-specification was planned to estimate three additional models. Two of these models (4 and 5) expand models 2 and 3 into bi-factor models by adding a general latent factor loading onto each subtest amongst the three latent ability factors (i.e., analytical, practical, and creative abilities) and the three latent domain factors (i.e., numerical, verbal, and figural item representation), respectively. Finally, model 6 specifies six correlated latent factors (three latent ability factors and three latent domain factors). It was hypothesized that introducing the additional factors would improve the model fit compared with models 1, 2, and 3.

It was planned to appraise the fit of three competing exploratory structural equation models (ESEM), models 7-9, to determine the best-fitting and most interpretable model. This plan was implemented because CFA implies restrictive assumptions on the measurement models, such as the absence of cross-loadings of a particular subtest on multiple abilities or domain factors. ESEM rose to prominence since Marsh and colleagues (2010) examined the fit of the 60-item NEO-Five Factor Inventory of personality factors. To our knowledge, ESEM has yet to be used to test the fit of a multi-ability, multi-domain measure of cognitive abilities such as Aurora-*a*. Because of their documented advantages (e.g., presumed better overall model fit, less biased factor loadings; Thöne et al., 2021) and because multidimensional items in psychological assessments are common (Sellbom & Tellegen, 2019), we aimed to estimate ESEMs to appraise the dimensionality of Aurora-*a*. More importantly, the underlying assumptions of ESEM are consistent with notions of the TSI that acknowledge the overlapping nature of analytical, practical, and creative abilities.

Consequently, ESEMs were estimated for abilities (analytical, practical, and creative thinking; model 7), domains (figural, numerical, verbal item presentation; model 8), and the combination of abilities and domains (model 9). It was hypothesized that the ESEM would yield a better fit to the data compared to the more restrictive CFA, in which the only freely estimated loadings are the ones of the latent factors that the subtests were designed to measure, whereas all other loadings are fixed to zero (commonly referred to as the independent cluster model, ICM). Gender differences were examined by testing the measurement invariance of the best-fitting model for female and male students. Four nested and increasingly restricted models (described below) were estimated and compared following the recommendations by Marsh and colleagues (Marsh et al., 2009; Morin et al., 2016). A previous analysis of mean differences found that females performed better than males on the five creativity subtests of Aurora-*a* (Mourgues, Hein, et al., 2016). However, none of the studies referenced above has examined measurement invariance of analytical, practical, and creative abilities by gender. Therefore, no specific hypotheses were formulated. Finally, the third aim was to evaluate evidence of concurrent criterion validity by estimating correlations of Aurora-*a* with external criterion measures of academic performance and cognitive ability.

Method

Participants

Data were collected from 3470 students in grades 4 (*n* = 419; 12.1%), 5 (*n* = 1359; 39.2%), and 6 (*n* = 1692; 48.8%) using a nonexperimental, observational design[1]. More than half of the participants identified as male (*n* = 1808; 52.1%). Students were recruited from 39 schools in the UK (*n* = 2693; 77.6%) and seven schools in the US (*n* = 777; 22.4%). Student-level

[1] Note that grades 4-6 are called "Years" in the UK, with Year 5 = US grade 4, Year 6 = US grade 5, and Year 7 = US grade 6.

information about race, ethnicity, and socioeconomic status was not available. The sample size was determined based on practical considerations and the availability of resource constraints (e.g., the time required to score open-ended responses). Data collection took place from 2006 through 2011. In the UK, all samples were convenience samples recruited from schools that had expressed interest in participating in this study. These schools were located in three cities in the north of England, in generally economically disadvantaged regions; the region's administration was interested in approaching their student body with an assessment that could capture a variety of human talent. In the US, sampling was conducted under similar conditions of interest. Approximately 85% of the total sample approached participated in the study. No incentives or payments were provided to participants. The Institutional Review Board at Yale University approved the study.

Assessments

Aurora-*a*. The 17 subtests of the Aurora Battery that assess a child’s analytical/memory, creative, and practical abilities within the figural, numerical, and verbal domains were designed to augment conventional IQ assessment, therefore it is called Aurora-*a* (Chart et al., 2008). The subtests capture students’ performance in three general abilities (analytical, practical, creative) and three functional domains (figural, numerical, and verbal). They also varied by response type, including multiple-choice, fill-in-the-blank, short answer (such as a single number), and open-ended free responses. Such varied formats were proposed to provide students with multiple modalities to demonstrate several types of skills. There are six analytical subtests: two in the figural domain (*Shapes*, 10 items; *Floating Boats*, 10 items), two in the verbal domain (*Homophone Blanks*, 20 items; *Metaphors*, 9 items), and two in the numerical domain (*Letter Math*, 5 items; *Algebra*, 5 items). There are six practical subtests: two in the figural domain

(*Paper Cutting*, 10 items; *Toy Shadows,* 8 items), two in the verbal domain (*Silly Headlines*, 11 items; *Decisions*, 3 items), and two in the numerical domain (*Maps*, 10 items; *Money Exchange*, 5 items). There are five creative subtests: two in the figural domain (in which visuals serve as the stimuli for creativity; *Book Covers*, 5 items; *Multiple Uses*, 5 items), two in the verbal domain (*Inanimate Conversations*, 10 items; *Figurative Language*, 12 items), and one in the numerical domain (*Cartoon Numbers*, 7 items). Of these five creative subtests, four are open-ended. Among the analytical subtests, one elicits open-ended responses (*Metaphors*). Two raters scored all open-ended responses to estimate inter-rater reliability (described below). Item score averages across raters were used in the analyses.

External Criterion Measures. Several academic achievement and cognitive ability tests were administered to collect data for assessment validation. Detailed information on these measures and their psychometric properties is available in the supplemental materials. The Cognitive Abilities Test (CogAT; Lohman, 2012) was administered to the US sample to appraise students' learned verbal, quantitative, and non-verbal reasoning abilities used in all areas of academic experiences rather than IQ. The Illinois Standards Achievement Test (ISAT) evaluates achievement in reading, writing, and mathematics in grades 3-8 and science in grade 4. The ISAT was administered to the US sample. The General Certificate Secondary Education (GCSE) is a level of general qualification that can apply to a wide range of curriculum areas taken in the UK by secondary school students aged 16, with English, math, and science being compulsory subjects. The Middle Years Information System (MidYIS; Centre for Evaluation and Monitoring, 2010) is a 45-minute baseline assessment for students entering secondary schools in the UK, with subtests in the areas of mathematics, vocabulary, skills (proofreading and perceptual speed and accuracy), and non-verbal reasoning (Tymms & Coe, 2003). The Year 11

Information System (YELLIS) is another progress monitoring test designed by the CEM commonly used for students aged 14-16 years in UK schools comprising mathematics, vocabulary, and perceptual/pattern reasoning sub-tests (Tymms & Coe, 2003).

Procedure

Paper copies of the Aurora Battery were used to collect the data. For the UK Year 7 administration, the 17 subtests were bundled into 3 booklets (A, B, C) with 8-9 subtests per booklet. The completion of each booklet took, on average, about 60 minutes. Data were collected in three sittings on three separate days, not necessarily consecutive, but generally within a single week. For the Year 5 and 6 students in the UK and most US administrations, the tests were bundled into six booklets (A-F), with 4-5 subtests per booklet. Booklets were administered in 45-minute blocks, with two blocks administered daily with a 15-minute break in between, during which children were allowed to get up and stretch their legs. Testing was completed in three days, not necessarily consecutive, but generally within a single week. All booklets in the data collections were gathered in two parallel forms, in which the order of subtests was reversed to address the order effect and increase the probability of gathering data from all subtests if students ran out of time. For the US-based Aurora data collection, informed consent processes obtaining written consent were used. For the UK-based data collection, all school principals in the school district agreed to participate in the study. Principals were asked to write to all the parents in their schools stating that they had agreed to assist with the trialing of an educational test that was being introduced across the school district. Testing would be conducted in class groups rather than individually. It was communicated to parents that the obtained results would inform the further design and development of the test and would not be used to make educational judgments or decisions about the children’s schooling. They were further informed that should they prefer their child not to participate, their child would be given alternative classroom activities during the scheduled testing session.

Data collectors were trained in an hour-long session in which they were introduced to the battery, shown the administration design, and instructed on how to administer the battery to students. Ten raters scored the open-ended questions. The raters were first familiarized with the scoring rubric using a training dataset of 50 students. In the training exercise, discrepancies in ratings were discussed until verbal consensus was reached, then the 50 tests were re-scored until each item agreement level reached .70 (Spearman's rank-order correlation coefficient, r_s). Following the training procedures, each pair of raters scored the remaining data individually, overlapping on about 10% of the records to allow the calculation of inter-rater agreement. Scores are deemed sufficiently reliable if an agreement above $r_s = .70$ is maintained for each pair of raters. Overall, the interrater reliability was acceptable across the open-ended subtests for the UK and the US samples. For pairs of raters in the UK, r_s ranged between 0.77 and 0.87 for Book Covers, 0.80 to 0.89 for Multiple Uses, 0.81 to 0.85 for Cartoon Numbers, 0.86 to 0.95 for Conversations, and 0.83 to 0.84 for Metaphors. For the US, complete agreement data were unavailable. However, the paired-rater process and minimum value for r_s were the same, meaning all pairs of raters needed to reach the agreement at 70% before data processing.

Analyses

The analyses aimed to compare the fit of a series of alternative models. Before estimating the models, the extent and pattern of missing data and data distributions were evaluated. Several students had missing data on entire subtests. In particular, between 12% (*Shadows*) and 30.1% (*Algebra*) of the total sample had missing data on all items of a particular subtest. Potential reasons for this amount of missing data include subtest order (e.g., *Algebra* was last in one of the booklets and, therefore, more likely to be skipped, mainly if children ran out of time) and a higher likelihood of skipping subtests that presented math problems (e.g., *Algebra*).

Missing item-level data ranged from 1.96% (*Shadows*) to 43.65% (*Homophones*) for an average of 11.62% across subtests. Full information maximum likelihood (FIML) was used to handle missingness because it uses all available data to estimate model parameters. FIML preserves the data distribution and produces unbiased estimates and standard errors under the missing at-random assumption. For subtest models, either the variance of the latent variables or the loading of the first item was fixed to 1 to ensure model identification. All models were estimated using Mplus version 8.8 (Muthén & Muthén, 2022). The weighted least squares mean and variance adjusted (WLSMV) estimator was used for subtests with binary or ordered categorical manifest indicators. A robust maximum likelihood (MLR) estimator (Satorra & Bentler, 1994) was used for subtests with open-ended questions because items of these subtests reflect the average rating scores across multiple raters.

We first conducted separate CFAs for each subtest (17 models) using the respective subtest items to confirm subtest unidimensionality and estimate subtest reliability (McDonald's ω). Second, alternative models 1 through 9 were tested using CFA (models 1-6) and ESEM (models 7-9). Because of the exploratory nature of the ESEM, it was essential to develop an *a-priori* coding scheme to evaluate the performance of the 17 Aurora-*a* subtests. For this coding scheme, we devised three distinguishable categories: (a) exemplary subtests that loaded significantly[2], substantially (i.e., ≥ 0.32), and positively onto only the one latent variable (ability or domain) that they were designed to assess; (b) acceptable subtests that showed significant

[2] The threshold value of α was set to 0.05 to determine statistical significance. Alpha level correction was not applied because (1) the relatively large sample provides sufficient statistical power, and (2) the analyses focus on determining patterns of significant factor loadings and not on the rejection of a specific null hypothesis.

target loadings (i.e., loadings on the subtest that they were designed to assess) but also demonstrated significant non-target loadings (i.e., loadings on subtests that they were not designed to assess); and (c) problematic subtests with substantial non-target loadings in the absence of a significant and positive target loading.

It was planned to evaluate the acceptable subtests in four tiers, with the first tier reflecting the most acceptable subtests. In the first tier, we grouped subtests with substantial standardized target loadings (≥ 0.32) with only one or two significant but small non-target loadings. In the second tier, subtests showed substantial standardized target loadings (0.32) and one substantial non-target factor loading. The third tier of acceptable subtests comprises the subtests with significant but small target loadings with at least one substantial standardized non-target loading (≥ 0.32). Lastly, the fourth tier included subtests showing significant target factor loadings but no substantial loading on any latent variable and at least one significant non-target loading. Finally, problematic subtests showed no significant target factor loading and significant, substantial non-target factor loadings. To test measurement invariance for male and female students, we estimated four models (Marsh et al., 2009; Morin et al., 2016): configural invariance (i.e., the same number of factors and pattern of loadings in both groups), weak factorial/measurement invariance (i.e., factor loadings were constrained to be invariant), strong factorial/measurement invariance (i.e., factor loadings and item intercepts were constrained to be invariant), and latent mean invariance (i.e., factor loadings, item intercepts, and latent factor means were constrained to be invariant in both groups; factor means were constrained to zero for model identification).

Model Evaluation

Model fit was assessed using established criteria (Hu & Bentler, 1999; Marsh et al., 2005): the comparative fit index (CFI), the root-mean-square error of approximation (RMSEA), and standardized root mean square residual (SRMR). Values greater than 0.90 and 0.95 for the CFI reflect an acceptable and excellent fit to the data, respectively. RMSEA estimates of less than 0.05 and 0.08 indicate an excellent fit and an acceptable fit to the data, respectively. SRMR values below 0.05 indicate an acceptable fit to the data. Overall goodness-of-fit, model parsimony, and interpretability were the main criteria for selecting one model over another. In addition to the global model fit indices, local fit testing was examined by checking residual correlations (Kline, 2016). Multiple fit measures were used as criteria to appraise measurement invariance by gender. In conventional CFA models, less than a .010 decrease in CFI (Cheung & Rensvold, 2002) and less than a .015 increase in RMSEA (Chen, 2007) indicate support for measurement invariance of the more restrictive model when using continuous indicators. Because this study tested a series of ESEM and the sample size was large (i.e., more than 800 individuals per group), several fit measures (i.e., the Akaike information criterion, AIC; Bayesian information criterion, BIC; sample size adjusted BIC, corBIC) were used to appraise measurement invariance (Cao & Liang, 2021).

Transparency and Openness

This study was not preregistered. Data, materials, and code/syntax are available by contacting the corresponding author.

Results

Aim 1: Subtest Model Testing

First, the unidimensionality of each of the 17 Aurora-*a* subtests was tested using CFA. Model results (fit indices) and reliability estimates (McDonald’s ω) are reported in Table 1. Model fit was not assessed for the *Decisions* subtest as applying the scoring rubric resulted in a just-identified model with three manifest indicators. Two items of the *Shapes* subtest were removed due to negative factor loadings (items 6 and 8), and another two were removed due to a non-significant factor loading (items 7 and 9). One item of the *Paper Cutting* subtest (item 3) was removed due to a negative factor loading. The first item of the *Letter Math* subtest was removed due to a non-significant factor loading. Following these model modifications, global fit indices across all subtests were in the acceptable range and supported the unidimensional structure of all Aurora-*a* subtests. Factor scores were saved from these models for subsequent model evaluation.

Aim 2: Evaluation of Aurora-*a*'s Factor Structure

Based on the estimated factor scores, nine models were estimated to test the internal factor structure of Aurora-*a* as a whole (i.e., including abilities and domains in the same model). Table 2 summarizes the fit indices of the tested models. As further described in this section, an ESEM outperformed CFA models on overall goodness-of-fit, alignment with the underlying theory, model parsimony, and interpretability for abilities (analytical, practical, and creative) and domains (figural, numerical, and verbal). Parameter estimates and the corresponding standard errors are reported in Supplementary Tables 2-10. Models 1 through 6 were estimated using CFA. The general factor model (model 1) yielded a poor fit to the data. The model fit of models 2 (three ability factors) and 3 (three domain factors) was somewhat better than the general factor model, but the CFI values were unacceptable. Models 4, 5, and 6 yielded acceptable global fit indices, yet these models were disregarded for several reasons. First, in model 4, the correlation between analytical and practical abilities was estimated at 0.968, signifying a strong overlap and little empirical discrimination between these two conceptually different abilities.

Furthermore, in this model, the factor loadings of three of the six analytical ability subtests (*Homophones*, *Letter Math*, *Metaphors*) and one practical subtest (*Headlines*) were not statistically significant, demonstrating difficulties in pinpointing the underlying meaning of this latent variable. These findings also demonstrate that these four subtests are multi-faceted and mismodeled using the restrictive ICM-CFA that assumes a one-to-one correspondence of indicators (here: subtest scores) to latent variables; that is, no cross-loadings are included in the model. Model 6 was excluded for similar reasons, particularly because of intercorrelations between the latent ability variables greater than 1, indicating model misspecification. Model 5 yielded an acceptable fit to the data. However, the factor loadings of the respective subtests on the latent figural, verbal, and numerical domain variables were all negative and, therefore, of no relevance in addition to the variance accounted for by the general factor. Because none of the CFA models showed a satisfactory fit or an interpretable factor structure, we pursued the planned ESEM to purposefully allow for cross-loadings of subtests on latent variables other than the ones they were developed to measure. Specifically, model 7 represents an ESEM in which the loadings of subtests developed to measure one of the three abilities were freely estimated for the respective ability. The remaining factor loadings were estimated as well but modeled to approximate zero. Model 8 is an analogous ESEM for the three latent domain variables (i.e., figural, numerical, verbal). By definition, both models had the same degrees of freedom and showed equivalent model fit.

We first interpret the model parameters of model 7. The results suggested that subtests are multi-faceted and can be empirically clustered into groups of exemplary, acceptable, and problematic subtests based on the pattern of their factor loadings. Regarding the exemplary subtests, the *Multiple Uses* and *Shadows* subtests fell into this category. Regarding the acceptable subtests, five in the first tier had substantial standardized target loadings (≥ 0.32), with only one or two significant but small non-target loadings: *Book Covers, Cartoon Numbers,* *Paper Cutting*, *Homophones*, and *Conversations*. In the second tier, two subtests showed substantial standardized target loadings (≥ 0.32) and one substantial non-target factor loading: *Algebra* and *Money*. The third tier of acceptable subtests comprised the subtests with significant but small target loadings but with at least one substantial standardized non-target loading (≥ 0.32): *Decision Making*, *Figurative Language*, and *Headlines*. Lastly, in the fourth tier, three subtests showed significant target factor loadings but no substantial loading on any latent ability variable and at least one significant non-target loading: *Letter Math*, *Maps*, and *Metaphors*.

Finally, two subtests were problematic because they showed significant, substantial non-target factor loadings but no significant target factor loading: *Boats* and *Shapes*. In model 7, 15 of the 17 subtests were deemed acceptable, albeit to varying degrees, measuring the analytical, practical, and creative abilities posited by the underlying TSI. A sensitivity analysis was performed by removing both problematic subtests (*Boats* and *Shapes*; data not tabulated). As expected, removing both subtests led to several major changes in the factor solution, particularly regarding the interpretability of the factor loadings. For instance, *Conversations* emerged as a new exemplary subtest together with *Multiple Uses*, whereas *Metaphors* emerged as a problematic subtest. Thus, it was decided to retain all subtests in model 7 for two reasons. First, although the *Boats* and *Shapes* subtests were problematic in that they unexpectedly showed no significant loading on the latent variable measuring analytical abilities, variations in both subtests were significantly and substantially accounted for by practical abilities. Second, both subtests showed no significant non-target loadings and can contribute meaningful information to understanding individual differences in practical abilities in this sample. The correlations between the latent variables were .57 (SE = .028, *p* < .001) for analytical and creative abilities, .42 (SE = .038, *p* < .001) for analytical and practical abilities, and .46 (SE = .037, *p* < .001) for practical and creative abilities.

In the ESEM of domains (model 8), there were three exemplary subtests that are all based on verbal forms of item presentation: *Homophones*, *Metaphors*, and *Figurative Language*. There were seven subtests in the first tier of the acceptable subtests. Three of those subtests (*Conversations*, *Headlines*, *Decision Making*) are based on verbal forms of item presentation, whereas another four subtests (*Boats*, *Shapes*, *Shadows*, *Paper Cutting*) comprised items with figural representation. There were two subtests with numerical item presentation in the third tier of the acceptable subtests: *Algebra* and *Money*. In the fourth tier, the subtest *Multiple Uses* was another acceptable subtest with figural item presentation. Four subtests were classified as problematic: *Cartoon Numbers*, *Book Covers*, *Letter Math*, and *Maps*. Although *Cartoon Numbers* had a significant target loading on the numerical factor (which it was designed to measure), the loading was negative and small (less than -0.32). *Book Covers* had a significant target loading but loaded substantially and negatively on the numerical factor. Overall, the verbal domain was the only latent factor comprised of exemplary or acceptable subtests. The figural domain was well-represented by four acceptable subtests in the first tier and one acceptable subtest in the fourth tier, but it also included a problematic subtest (*Book Covers*). Finally, the numerical domain comprised more problematic (three) than acceptable (two) subtests. In this model, the verbal and figural domains correlated at $r = 0.72$ (SE = .045, $p < .001$). The numerical domain was not significantly related to the figural ($r = 0.21$, SE = .189, $p = .26$) and the verbal ($r = -0.07$, SE = .144, $p = .62$) domains. Regarding model 9, the parameter estimates revealed only non-significant loadings of subtests for the figural and numerical domains. Thus, this model was deemed less parsimonious than models 7 and 8 and was not selected for further analyses.

Concerning measurement invariance by gender, results showed that the latent means of analytical, practical, and creative abilities were comparable for male and female students. Specifically, the least restrictive model (configural invariance) fit the data well, χ^2 (176) = 334.36, CFI = .983, TLI = .974, RMSEA = .023, SRMR = .02, AIC = 51225.15, BIC = 52234.06, corBIC = 51712.96. The more restrictive weak factorial/measurement invariance model showed minor differences in model fit compared to the configural invariance model, ΔCFI = .003, ΔTLI = -.001, ΔRMSEA = -.001, ΔAIC = 3.86, ΔBIC = 262.24, ΔcorBIC = 128.79. There was support for the strong factorial/measurement invariance model compared to the weak factorial/measurement invariance model based on the change in fit measures, ΔCFI = .008, ΔTLI = .008, ΔRMSEA = .003, ΔAIC = 53.59, ΔBIC = 32.53, ΔcorBIC = 11.95. Finally, the difference in model fit between the latent mean invariance model and the strong factorial/measurement invariance model was less than the recommended cut-offs, ΔCFI = .002, ΔTLI = .002, ΔRMSEA = .001, ΔAIC = 24.17, ΔBIC = 5.71, ΔcorBIC = 15.25. Descriptive statistics for the latent means from the measurement invariance models are not presented because latent factor means are fixed to zero for model identification. Fit indices of all tested measurement invariance models are presented in Supplemental Tables 11 and 12.

Aim 3: Estimates of Criterion Validity

Correlations were computed between the Aurora-*a* ability and domain scores with performance tests in the following academic areas: (a) non-verbal reasoning, (b) verbal reasoning, (c) math, (d) reading, (e) writing, (f) science, (g) humanities, and (h) language arts. The external criterion measure scores were obtained from one or more sub-tests of the CogAT and ISAT in the US sample and MidYIS, GCSE, and YELLIS in the UK sample. We included sub-samples of students for whom these data were available and only included subtests if they had data on at least 80 cases (Bonett & Wright, 2000). Overall, weak to strong correlations (ranging from *r* = .20 to .72) with academic performance assessments corroborated the criterion validity of Aurora-*a*'s ability scores, with the direction and magnitude of the estimates in the expected range. Figure 1 shows the pattern of correlations by academic domain. The Aurora-*a* domains and abilities were highly correlated with performance tests of reading (range: *r* = .53, *p* < .001 for the figural domain to *r* = .72, *p* < .001 for practical ability). The analytical and practical abilities and the verbal domain were moderately correlated with performance tests in math, verbal reasoning, science, humanities, and non-verbal reasoning. Finally, the overall magnitude of the correlations was smaller with writing (range: *r* = .20, *p* = 0.08 for the numerical domain and *r* = .46, *p* < .001 for the verbal domain), and the lowest correlations in the areas of math (*r* = .39, *p* < .05) and science (*r* = .31, *p* < .05) were with creative ability.

Discussion

This study examined the psychometric properties of Aurora-*a*, a cognitive assessment of analytical, practical, and creative abilities when exercised in figural, numerical, and verbal domains in middle childhood and early adolescence. Using data from 3470 students (1808, or 52.1%, identified as male) from the UK and the US, we examined the dimensionality, reliability, and validity of Aurora-*a*. It is important to note that Aurora-*a* has been translated into several languages and applied in several countries, indicating Aurora's potential to be used in a variety of settings around the world. However, in this case, given the need for validated instruments that are not readily available in many settings, we focused on the two convenience samples where such instruments were accessible. With this study, we aimed to contribute to the intellectual discourse on how to best define and measure cognitive abilities beyond those contributing to the psychometric *g*. Therefore, measurement model respecification (i.e., exclusion items with negative and/or non-significant factor loadings) was necessary to test the unidimensionality of the 17 subtests. We established support for our hypothesis that ESEM would outperform the more restrictive CFA used in previous studies on Aurora-*a*. Here, the CFA demonstrated several non-significant subtest factor loadings, factor intercorrelations greater than one, and negative factor loadings. The ESEM, in contrast, yielded an adequate fit to the data. By employing this analytic technique, we acknowledge that developing an assessment that fully differentiates specific cognitive abilities is challenging. Therefore, the findings presented here exemplify that, in some instances, CFA may not be well-suited to capture individual differences in students' performance on a set of theoretically related yet distinct cognitive abilities.

The studies that examined the factor structure of translated versions of Aurora-*a* in other countries relied on CFA (Aghababaei et al., 2016; Aslan & Soysal, 2021; Gubbels et al., 2016). This difference in analytic approaches and the fact that previous studies have typically excluded at least the *Shapes* subtest (Aghababaei et al., 2016; Gubbels et al., 2016) preclude a more thorough comparison of the similarities and differences in parameter estimates between the present study and the published research on Aurora. Despite these differences, several important conclusions about the performance of Aurora-*a* can be drawn across these studies.

First, the findings from the ESEM illustrate the multifaceted nature of several Aurora-*a* subtests. For instance, the ESEM of Aurora-*a* ability factors revealed several subtests that showed substantial (≥ 0.32) standardized target loadings on the factor they were designed to quantify but also at least one loading on a factor they were not designed to assess. These subtests include, for instance, three of the creativity subtests: *Book Covers, Cartoon Numbers,* and *Conversations.* Reliable variance in children's scores on the *Conversations* subtest was mostly accounted for by creative ability with additional contributions of analytical ability. This observation partly sheds light on children's underlying thought processes in response to the presented items. To generate dialogues between everyday objects, one may need to infer their (physical) features based on one’s knowledge about said objects and their relation to each other. This finding supports the notion that analytical and creative thinking, at least as conceived here, were linked, which has received a fair amount of attention in the past decade (e.g., Gerwig et al., 2021; Silvia, 2015). *Paper Cutting* is another example that demonstrates the interplay of multiple abilities. The findings showed that practical thinking accounts for most of the reliable variance compared to the other abilities. Analytical and creative thinking is also required to solve the items, albeit to a lesser degree than practical thinking. Some analytical abilities, such as mental rotation, are undoubtedly needed to infer the proper position of an unfolded paper. Second, the interpretation of the ESEM showed that variation in Aurora-*a* subtest scores related somewhat more consistently to abilities rather than the form of item presentation (i.e., figural, numerical, and verbal domains).

This study corroborates previous findings on the medium-level convergence between standardized achievement tests and Aurora-*a* scores. For example, Prieto and colleagues (2015) found that the verbal domain was related to a traditional measure of IQ in their sample of Spanish students. In contrast, the relationship was weaker with the figurative domain. In addition, Mandelman and colleagues (2013) administered a standardized achievement test covering reading, language, and math and observed higher correlations with practical (.62-.71) and analytical abilities (.55-.62) than with creative ability (.30-.37). With regards to science, the correlations with creative and/or analytical and practical abilities were generally similar to previous research in the UK (Kornilov et al., 2011; Mourgues, Tan, et al., 2016).

The present study also showed some important variations in the pattern of correlations due to the exploratory nature of the analyses. For example, Mandelman and colleagues (2013) observed the highest correlation to be between practical ability and math (.71). In contrast, in the present study, the highest correlation was between practical ability and reading (.72). In the present study, the Aurora-*a* numerical domain was positively correlated with math performance tests (.56), the verbal domain correlated with the verbal academic performance tests (.53), and the figural domain correlated with the non-verbal academic performance tests (.44). However, both numerical and verbal domains (~.65), as well as the figural domain (.53), and the three abilities (.59-.72), were highly correlated with the performance indicators of reading, likely because many Aurora-*a* tests have reading requirements. For example, the *Money* and *Algebra* subtests consist of word problems or math questions written as one or more sentences. Similarly, numerical subtests such as *Cartoon Numbers* have a verbal component (students are required to write a short paragraph about two numbers; Tan et al., 2013), as do practical subtests such as *Decisions*. Previous studies have also found small correlations between recreational reading and the three ability scores, but not with other recreational activities such as playing outdoors or listening to music (Hein et al., 2014). In the present study, the correlation between creativity and reading (.59) was similar to previous studies (r = .35-.53; Kornilov et al., 2011; Mourgues, Tan, et al., 2016), but the correlation between creative ability and writing was lower in the present study (.29) than in previous studies (r = .42-.53; Kornilov et al., 2011; Mourgues, Tan, et al., 2016) possibly due to the cross-loadings of the open-ended items on the analytic and practical abilities. It is possible that the sample of the present study exhibited a greater range of performance in reading comprehension and written expression and included children who experienced difficulty in reading the items of the practical tests and/or providing accurate written responses[3]. Future studies are needed to comprehensively evaluate the relation between reading and/or writing (dis)ability and performance on Aurora-*a* subtests.

Integration with Developmental Frameworks of Intelligence

Although models incorporating a general factor were not selected as the best models in the present study, the pattern of results from these models still holds significant implications for developmental models of intelligence, such as the differential-developmental theory (Demetriou et al., 2018; Demetriou et al., 2022; Demetriou et al., 2023) and the differentiation hypotheses of intelligence (Breit et al., 2022; Breit et al., 2021; Breit et al., 2024). The findings from Model 4, which incorporates a general factor (g) and specific factors for analytic, practical, and creative thinking, align closely with findings from Demetriou and colleagues (2022), particularly regarding reasoning. Demetriou and colleagues found that reasoning skills become more distinct and specialized over time, contributing significantly to cognitive performance independently of general intelligence. This involves complex processes such as logical thinking, problem-solving, and integrating information, which align with the specific factor loadings for analytic tasks in Model 4, such as *Figures* (0.403). High loadings on the general factor for tasks like *Algebra* (0.604) also emphasize the role of general intelligence, supporting the dual influence of general and specific cognitive skills.

On the other hand, Model 5 results demonstrate that domain-specific abilities have stronger loadings compared to general intelligence (g), aligning with Breit and colleagues (2021), who emphasize that numerical and verbal reasoning become more specialized over time, diluting their correlations with the *g*-factor. In Model 5, numerical tasks, such as *Algebra* (-0.66)

[3] Coefficients of variation cannot be directly compared because previous studies did not report the mean and standard deviations on the academic assessments or did not report them on a scale that would allow meaningful comparisons.

and *Letter Math* (-0.368), show high loadings on domain-specific factors, indicating these abilities are more influenced by specific skills than by g. Similarly, higher loadings on domain-specific factors in figural (e.g., *Paper Cutting* = -0.68; *Shadows* = -0.546) and verbal domains (e.g., *Decision Making* = -0.529; *Figurative Language* = -0.574) underscore the significance of specialized cognitive abilities. However, the presence of negative loadings on specific factors and positive loadings on the general factor suggests that some tasks are influenced by overall cognitive ability, indicating a reliance on general intelligence for performance. This pattern reflects the complexity of these tasks, which draw on multiple cognitive processes, leading to a stronger association with general intelligence. These findings corroborate Breit's observations that general intelligence becomes less dominant as specific abilities differentiate and develop. Overall, our results highlight the importance of focusing on domain-specific skills in educational practices and cognitive assessments to capture the nuanced development of cognitive abilities. Future research should investigate developmental trajectories in detail to further understand the dynamics of cognitive specialization.

Constraints on Generality

The study sample includes students from 4th to 6th grades in the US and the UK. Although efforts were made to increase the diversity of this sample, unexpected difficulties arose when we encountered low reading levels in English and large numbers of non-English speaking children. Our existing form (in English only) was a limitation. Because of the convenience sampling procedure, the sample was not sufficiently diverse or representative for the study findings to generalize to the broader population, including younger or older students, students in other countries, neurodiverse individuals, and students who identify outside the gender binary. There is a need to replicate the study's findings by directly testing the identified factor models in a more diverse sample. Aurora-*a* is a paper-pencil assessment that requires students to read and respond to multiple-choice, fill-in-the-blank, short-answer, and open-ended items. These critical features of the assessment must be maintained to measure cognitive abilities in a replication attempt.

Regarding the study procedures, Aurora-*a* is an individual assessment using printed copies administered in a group (e.g., classroom) setting. Anyone who has completed the appropriate one-hour training should be able to administer the assessment. However, given the substantial amount of missing data in some of the subtests in international samples, more emphasis must be placed on ensuring that students understand the instructions, have sufficient time, and are comfortable responding to the items. Furthermore, data used for this study were collected 10 to 15 years ago. We can only speculate about changes in the historical context in the places where data were collected. In the past decade, in addition to (meta-)cognitive, social, and emotional skills, the promotion of students' practical skills has indeed received attention as described, for instance, in the OECD's "Skills for 2030" framework (OECD, 2019). It is reasonable to expect that the ways in which students are supported in strengthening their abilities (as indexed by Aurora-*a*) are constantly changing over time. Therefore, future studies should attempt to better capture the contextual factors that may impact students' performance on Aurora-*a*'s items and their familiarity with the test format. The present analyses were also limited because fewer participants were administered the external criterion measures (e.g., *n* = 1,298 for math and *n* = 128 for reading). More research on the validity of Aurora-*a* is warranted with samples that are heterogeneous in age, socio-demographic composition, and literacy skills.

Implications and Conclusions

Although this study did not aim to empirically test the TSI (Sternberg, 2020), our findings support some of its tenets and provide a foundation for testing a range of cognitive abilities necessary for students' global functioning (in and out of school). We did not include students' performance on facets of psychometric *g* because the evidence base for Aurora-*a* supports the multi-dimensionality of intelligence. However, we refrain from positioning the findings in favor of a particular theoretical notion of intelligence. Instead, we encourage studies that aim to better understand profiles of cognitive abilities in middle childhood and early adolescence that consider a broad and diverse spectrum of human cognitive abilities. Aurora-*a* can play a crucial role in such studies based on the psychometric evidence presented here. The battery aims to augment the identification of gifted children across the assessed areas. By focusing on analytical, practical, and creative abilities, Aurora-*a* brings attention to the unconventional range of abilities and skills that children come to school with that may support the pursuit of their goals and well-being. Aurora-*a* recognizes the need to foster students' creative skills and provides an example of how to do so. While assessments for practical skills are challenging, our qualified effort in tackling this problem remains the only attempt to do so with this age group to date. Conventional tests of abilities have played a significant role in responding to societal challenges of the 20th century. The 21st century has introduced a new spectrum of challenges that call for various other abilities needed to address these challenges (UNESCO, 2023). Aurora-*a* is an assessment that can help identify and foster these abilities.

Reference List

Aghababaei, S., Malekpour, M., Kajbaf, B., & Abedi, A. (2016). Confirmatory Factor Analysis of Aurora-a Battery on Children. *Modern Applied Science*, *10*(10), 99-105. https://doi.org/10.5539/mas.v10n10p99

Aljughaiman, A. M., & Ayoub, A. E. A. (2012). The Effect of an Enrichment Program on Developing Analytical, Creative, and Practical Abilities of Elementary Gifted Students. *Journal for the Education of the Gifted*, *35*(2), 153-174. https://doi.org/10.1177/0162353212440616

Aslan, A. E., & Soysal, S. (2021). Reliability and Validity of the Turkish Version of Aurora-a Intelligence Test Battery. *Education Quarterly Reviews*, *4*(2), 214-225. https://doi.org/10.31014/aior.1993.04.02.241

Ayoub, A. E. A., & Aljughaiman, A. M. (2016). A predictive structural model for gifted students' performance: A study based on intelligence and its implicit theories. *Learning and Individual Differences*, *51*, 11-18. https://doi.org/10.1016/j.lindif.2016.08.018

Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika*, *65*(1), 23-28.

Breit, M., Brunner, M., Molenaar, D., & Preckel, F. (2022). Differentiation hypotheses of intelligence: A systematic review of the empirical evidence and an agenda for future research. *Psychological Bulletin*, *148*(7-8), 518-554. https://doi.org/10.1037/bul0000379

Breit, M., Brunner, M., & Preckel, F. (2021). Age and ability differentiation in children: A review and empirical investigation. *Dev Psychol*, *57*(3), 325-346. https://doi.org/10.1037/dev0001147

Breit, M., Scherrer, V., Tucker-Drob, E. M., & Preckel, F. (2024). The stability of cognitive abilities: A meta-analytic review of longitudinal studies. *Psychol Bull*, *150*(4), 399-439. https://doi.org/10.1037/bul0000425

Cao, C., & Liang, X. (2021). Sensitivity of Fit Measures to Lack of Measurement Invariance in Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(2), 248-258. https://doi.org/10.1080/10705511.2021.1975287

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

Centre for Evaluation and Monitoring. (2010). *MidYIS assessments*. https://www.cem.org/midyis

Chart, H. E., Grigorenko, E. L., & Sternberg, R. J. (2008). Identification: The Aurora Battery. In J. A. Plucker & C. M. Callahan (Eds.), *Critical issues and practices in gifted education* (pp. 345-365). Prufrock Press.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464-504. https://doi.org/https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255. https://doi.org/https://doi.org/10.1207/S15328007SEM0902_5

CTB/McGraw-Hill. (2010). *TerraNova*.

Demetriou, A., Makris, N., Spanoudis, G., Kazi, S., Shayer, M., & Kazali, E. (2018). Mapping the Dimensions of General Intelligence: An Integrated Differential-Developmental Theory. *Human Development*, *61*(1), 4-42. https://doi.org/10.1159/000484450

Demetriou, A., Mougi, A., Spanoudis, G., & Makris, N. (2022). Changing developmental priorities between executive functions, working memory, and reasoning in the formation of g from 6 to 12 years. *Intelligence*, *90*. https://doi.org/10.1016/j.intell.2021.101602

Demetriou, A., Spanoudis, G., Christou, C., Greiff, S., Makris, N., Vainikainen, M. P., Golino, H., & Gonida, E. (2023). Cognitive and personality predictors of school performance from preschool to secondary school: An overarching model. *Psychol Rev*, *130*(2), 480-512. https://doi.org/10.1037/rev0000399

Ellis, B. J., Abrams, L. S., Masten, A. S., Sternberg, R. J., Tottenham, N., & Frankenhuis, W. E. (2022). Hidden talents in harsh environments. *Development and Psychopathology*, *34*(1), 95-113.

Ferrando, M., Ferrándiz, C., Llor, L., & Sainz, M. (2016). Successful intelligence and giftedness: an empirical study. *Anales de Psicología*, *32*(3), 672-682. https://doi.org/10.6018/analesps.32.3.259431

Gardner, H. (2011). *Frames of mind: The theory of multiple intelligences*. Basic Books.

Gardner, H. E. (2006). *Multiple intelligences: New horizons*. Basic Books.

Gerwig, A., Miroshnik, K., Forthmann, B., Benedek, M., Karwowski, M., & Holling, H. (2021). The Relationship between Intelligence and Divergent Thinking-A Meta-Analytic Update. *Journal of Intelligence*, *9*(2). https://doi.org/10.3390/jintelligence9020023

Grigorenko, E. L., Jarvin, L., & Sternberg, R. J. (2002). School–based tests of the triarchic theory of intelligence: Three settings, three samples, three syllabi. *Contemporary Educational Psychology*, *27*, 167–208.

Grigorenko, E. L., Meier, E., Lipka, J., Mohatt, G., Yanez, E., & Sternberg, R. J. (2004). Academic and practical intelligence: A case study of the Yup'ik in Alaska. *Learning and Individual Differences*, *14*, 183–207. https://doi.org/10.1016/j.lindif.2004.02.002

Grigorenko, E. L., & Sternberg, R. J. (2001). Analytical, creative, and practical intelligence as predictors of self-reported adaptive functioning: A case study in Russia. *Intelligence*, *28*, 1–17.

Gubbels, J., Segers, E., Keuning, J., & Verhoeven, L. (2016). The Aurora-a Battery as an Assessment of Triarchic Intellectual Abilities in Upper Primary Grades. *Gifted Child Quarterly*, *60*(3), 226-238. https://doi.org/10.1177/0016986216645406

Hedlund, J., Wilt, J. M., Nebel, K. R., Ashford, S. J., & Sternberg, R. J. (2006). Assessing practical intelligence in business school admissions: A supplement to the graduate management admissions test. *Learning and Individual Differences*, *16*, 101-127.

Hein, S., Tan, M., Aljughaiman, A., & Grigorenko, E. L. (2014). Characteristics of the home context for the nurturing of gifted children in Saudi Arabia. *High Ability Studies*, *25*, 23-33.

Hodges, J., Tay, J., Maeda, Y., & Gentry, M. (2018). A meta-analysis of gifted and talented identification practices. *Gifted Child Quarterly*, *62*(2), 147-174.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kornilov, S. A., Tan, M., Elliott, J. G., Sternberg, R. J., & Grigorenko, E. L. (2011). Gifted identification with Aurora: Widening the spotlight. *Journal of Psychoeducational Assessment*, *30*(1), 117-133. https://doi.org/10.1177/0734282911428199

Lohman, D. F. (2012). *CogAT score interpretation guide*. Riverside.

Mandelman, S. D., Barbot, B., & Grigorenko, E. L. (2016). Predicting academic performance and trajectories from a measure of successful intelligence. *Learning and Individual Differences*, *51*, 387-393. https://doi.org/10.1016/j.lindif.2015.02.003

Mandelman, S. D., Barbot, B., Tan, M., & Grigorenko, E. L. (2013). Addressing the 'quite crisis': Gifted identification with Aurora. *Educational & Child Psychology*, *30*(2), 101-109.

Mandelmann, S. D., & Grigorenko, E. L. (2013). Questionning the unquestionnable: reviewing evidence for the efficacy of gifted education. *Talent Development & Excellence*, *5*(1), 125-137.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modelling. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275-340). Erlbaum.

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*(3), 471-491. https://doi.org/https://doi.org/10.1037/a0019227

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students' Evaluations of University Teaching. *Structural Equation*

Modeling: A Multidisciplinary Journal, *16*(3), 439-476. https://doi.org/10.1080/10705510903008220

Morin, A. J. S., Arens, A., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling*, *24*(1), 116-139. https://doi.org/10.1080/10705511.2014.961800

Mourgues, C., Hein, S., Tan, M., Diffley III, R., & Grigorenko, E. L. (2016). The role of non-cognitive factors in predicting academic trajectories of high school students in a selective private school. *European Journal of Psychological Assessment*, *32*, 84-94.

Mourgues, C., Tan, M., Hein, S., Elliott, J. G., & Grigorenko, E. L. (2016). Using creativity to predict future academic performance: An application of Aurora's five subtests for creativity. *Learning and Individual Differences*, *51*, 378-386. https://doi.org/10.1016/j.lindif.2016.02.001

Muthén, L. K., & Muthén, B. O. (2022). *Mplus Version 8.8*. In Muthén & Muthén.

OECD. (2019). *OECD Future of Education and Skills 2023 - Conceptual learning framework Skills for 2030*.

Prieto, D., Ferrándiz, C., Ferrando, M., & Bermejo, M. R. (2015). La Batería Aurora: una nueva evaluación de la inteligencia exitosa [Aurora Battery: A new assessment of successful intelligence]. *Revista de Educación*, *368*, 132-157. https://doi.org/10.4438/1988-592X-RE-2015-368-294

Ricciardi, C., Haag-Wolf, A., & Winsler, A. (2020). Factors associated with gifted identification for ethinically diverse children in poverty. *Gifted Child Quarterly*, *64*, 243-258.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance sturcture analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications to developmental research*. SAGE.

Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychol Assess*, *31*(12), 1428-1441. https://doi.org/10.1037/pas0000623

Silvia, P. J. (2015). Intelligence and Creativity Are Pretty Similar After All. *Educational Psychology Review*, *27*(4), 599-606. https://doi.org/10.1007/s10648-015-9299-1

Spearman, C. (1904). "General Intelligence": Objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201-292.

Speirs Neumeister, K. L., Adams, C. M., Pierce, R., Cassady, J. C., & Dixon, F. (2007). Fourth-grade teachers' perceptions of giftedness: Implications for identifying and serving diverse gifted students. *Journal for the Education of the Gifted*, *30*(4), 479-499. https://doi.org/https://doi.org/10.4219/jeg-2007-503

Stemler, S. E., Grigorenko, E. L., Jarvin, L., & Sternberg, R. J. (2006). Using the theory of successful intelligence as a basis for augmenting AP exams in Psychology and Statistics. *Contemporary Educational Psychology*, *31*(3), 344-376. https://doi.org/10.1016/j.cedpsych.2005.11.001

Stemler, S. E., Sternberg, R. J., Grigorenko, E. L., Jarvin, L., & Sharpes, K. (2009). Using the theory of successful intelligence as a framework for developing assessments in AP physics. *Contemporary Educational Psychology*, *34*(3), 195-209.

Sternberg, R. J. (2012). The triarchic theory of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment. Theories, tests, and issues.* (pp. 156-177). The Guilford Press.

Sternberg, R. J. (2020). The augmented theory of successful intelligence. In R. J. Sternberg (Ed.), *Cambridge handbook of intelligence* (2nd ed., Vol. 2, pp. 679-708). Cambridge University Press.

Sternberg, R. J., Castejón, J. L., Prieto, M. D., Hautamäki, J., & Grigorenko, E. L. (2001). Confirmatory factor analysis of the Sternberg triarchic abilities test in three international samples: An empirical test of the triarchic theory of intelligence. *European Journal of Psychological Assessment*, *17*, 1-16.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J., Snook, S., Williams, W. M., Wagner, R. K., & Grigorenko, E. L. (2000). *Practical intelligence in everyday life*. Cambridge University Press.

Sternberg, R. J., & Grigorenko, E. L. (2000). *Teachingfor successful intelligence*. Skylight Training and Publishing.

Sternberg, R. J., & Grigorenko, E. L. (2002). The theory of successful intelligence as a basis for gifted education. *Gifted Child Quarterly*, *46*(4), 265-277.

Sternberg, R. J., & Grigorenko, E. L. (Eds.). (2012). *The psychology of abilities, competencies, and expertise*. Cambridge University Press.

Sternberg, R. J., & The Rainbow Project Collaborators. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical and creative skills. *Intelligence*, *34*(4), 321-350.

Tan, M., Aljughaiman, A. M., Elliott, J. G., Kornilov, S. A., Ferrando-Prieto, M., Bolden, D. S., Adams-Shearer, K., Chart, H. M., Newman, T., Jarvin, L., Sternberg, R. J., & Grigorenko, E. L. (2009). Considering language, culture, and cognitive abilities: The international translation and adaptation of the Aurora Assessment Battery. In E. L. Grigorenko (Ed.), *Multicultural Psychoeducational Assessment* (pp. 443-468). Springer.

Tan, M., Barbot, B., Mourgues, C., & Grigorenko, E. L. (2013). Measuring metaphors: Concreteness and similarity in metaphor comprehension and gifted identification. *Educational and Child Psychology Journal*, *30*(2), 89-100.

Tan, M., Kilani, H., Markov, I., Hein, S., & Grigorenko, E. L. (2023). Assessing cognitive skills in early childhood education using a Bilingual Early Language Learner Assessment Tool. *Journal of Intelligence*, *11*(7), 143.

Tan, M., Mourgues, C. V., Aljughaiman, A., Ayoub, A., Mandelman, S. D., Zbainos, D., & Grigorenko, E. L. (2012). What the shadow knows: Assessing aspects of practical intelligence with Aurora's Toy Shadows. In H. Stoeger, A. Aljughaiman, & B. Harder (Eds.), *Talent development and excellence*. LIT.

Thöne, A.-K., Junghänel, M., Görtz-Dorten, A., Dose, C., Hautmann, C., Jendreizik, L. T., Treier, A.-K., Vetter, P., von Wirth, E., Banaschewski, T., Becker, K., Brandeis, D., Dürrwächter, U., Geissler, J., Hebebrand, J., Hohmann, S., Holtmann, M., Huss, M., Jans, T., . . . Döpfner, M. (2021). Disentangling symptoms of externalizing disorders in children using multiple measures and informants. *Psychological Assessment, 33*(11), 1065-1079.

Tymms, P., & Coe, R. (2003). Celebration of the success of distributed research with schools: the cem centre, Durham. *British Educational Research Journal*, *29*(5), 639-667. https://doi.org/10.1080/0141192032000133686

UNESCO. (2023). *Global initiative around assessment of 21st century skills*. Retrieved May 24, 2023 from https://www.unesco.org/en/articles/global-initiative-around-assessment-21st-century-skills

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842-861.

Table 1

Model Fit Indices for the 17 Aurora-a Subtests

	n	Number of items	χ^2 (*df*), *p*	RMSEA (90%-CI)	CFI	SRMR	ω
Analytical ability							
Boats	3020	10	433.30 (35), < .001	.061 (.056, .067)	.959	.056	0.88
Shapes	2454	10 / 6 [c]	12.67 (9), .178	.013 (.000, .028)	.987	.022	0.48
Homophones	2576	20	568.34 (170), < .001	.030 (.027, .033)	.994	.051	0.98
Metaphors [a]	2557	9	123.60 (27), < .001	.037 (.031, .044)	.958	.030	0.76
Letter Math	2758	5 / 4 [d]	28.54 (2), < .001	.069 (.048, .093)	.965	.059	0.80
Algebra	2424	5	15.76 (5), .008	.030 (.014, .047)	.988	.029	0.76
Practical ability							
Paper Cutting	3049	10 / 9 [e]	60.04 (27), < .001	.020 (.013, .027)	.987	.027	0.75
Toy Shadows	3054	8	73.34 (20), < .001	.030 (.022, .037)	.981	.033	0.74
Silly Headlines	2733	11	248.40 (44), < .001	.041 (.036, .046)	.995	.024	0.95
Decisions [a,b]	2871	3					0.62

Mapping	2809	10	711.30 (35), < .001	.083 (.078, .088)	.921	.055	0.86
Money Exchange	2873	5	25.31 (5), < .001	.038 (.024, .053)	.991	.021	0.79
Creative ability							
Book Covers [a]	2435	5	77.46 (5), < .001	.077 (.063, .093)	.974	.025	0.88
Multiple Uses [a]	2715	5	40.66 (5), < .001	.051 (.037, .066)	.987	.019	0.82
Conversations [a]	2706	10	123.65 (35), < .001	.031 (.025, .037)	.988	.019	0.89
Figurative Language	2551	12	116.93 (54), < .001	.021 (.016, .027)	.995	.027	0.90
Cartoon Numbers [a]	2973	7	150.56 (14), < .001	.057 (.049, .066)	.940	.034	0.72

Notes. *N* = 3470. [a] Robust maximum likelihood (MLR) estimator was used. [b] Model fit not assessed for just-identified models with zero degrees of freedom. [c] Four items of the *Shapes* subtest were removed due to negative factor loadings (items 6 and 8) or non-significant factor loadings (items 7 and 9). [d] One item of the *Letter Math* subtest (item 1) was removed due to a non-significant factor loading. [e] One item of the *Paper Cutting* subtest (item 3) was removed due to a negative factor loading.

Table 2

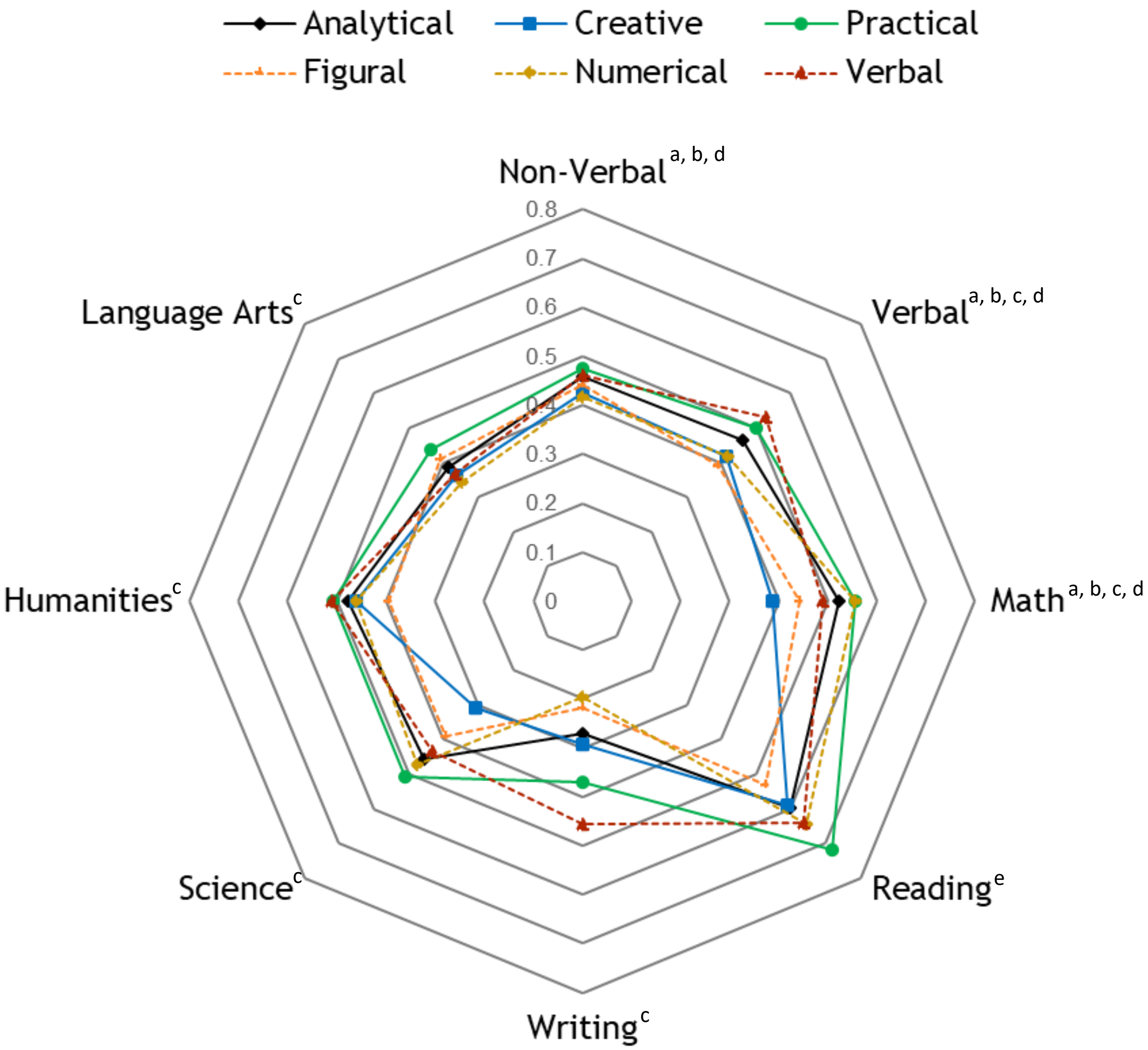
Model Fit Indices of Alternative Models

Model	χ^2 (*df*), *p*	RMSEA (90%-CI)	CFI	SRMR	AIC	Sample size adjusted BIC
1. General factor	1094.19 (119), < .001	.049 (.046, .051)	0.895	.045	52122.90	52274.60
2. Three abilities	950.02 (116), < .001	.046 (.043, .048)	0.910	.041	51981.79	52142.41
3. Three domains	836.33 (116), < .001	.042 (.040, .045)	0.922	.044	51865.37	52025.99
4. Three abilities and general factor	327.85 (99), < .001	.026 (.023, .029)	0.975	.022	51379.91	51591.10
5. Three domains and general factor	219.36 (99), < .001	.019 (.015, .022)	0.987	.017	51273.98	51485.17
6. Three abilities and three domains	218.48 (96), < .001	.019 (.016, .023)	0.987	.017	51277.91	51498.02
7. ESEM of abilities	250.74 (88), < .001	.023 (.020, .026)	0.982	.017	51319.47	51563.38
8. ESEM of domains	250.74 (88), < .001	.023 (.020, .026)	0.982	.017	51319.47	51563.38
9. ESEM abilities and domains	111.60 (49), < .001	.019 (.014, .024)	0.993	.010	51233.27	51593.18

Notes. *N* = 3470. All models were estimated using a robust maximum likelihood (MLR; Satorra & Bentler, 1994) estimator.

Figure 1

Correlations Between Aurora-a Scores and External Criterion Measures (Academic Performance and Cognitive Ability)



Note. [a]MidYIS, [b]YELLIS, [c]GCSE, [d]CogAT, [e]ISAT. Science scores were averaged for the science and additional science sub-tests of the GCSE. Humanities scores were averaged for the history and religious studies sub-tests of the GCSE. Language Arts scores were averaged for English literature, Design Technology, and Arts Design sub-tests of the GCSE. Sample sizes: Non-Verbal = 1150; Verbal = 1414; Math = 1298; Reading = 128; Writing = 81; Science = 296; Humanities = 438; Language Arts = 187. Correlations above 0.20 are significant at $p < 0.05$.