Declarations

Funding

This study was funded by the Flemish government.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics Approval and Consent

The present study is part of "The Columbus Project: a tool for study orientation towards higher education", to which the Ethical Commission of the Faculty of Psychology and Educational Sciences at Ghent University has given approval with reference to 2016/82. However, ethical approval was deemed unnecessary for this study involving human participants, as it adhered to local legislative and institutional requirements. The study utilized datasets derived from the Columbus project, commissioned by the Flemish Ministry of Education and Training. All participants provided written informed consent. Prior to registration, students agreed to terms and conditions developed in collaboration with the Data Protection Officer of the Ministry of Education and Training, aligning with the General Data Protection Regulation (GDPR). The Ministry rigorously adheres to GDPR principles in the handling of personal data, and any sharing of such data necessitates adherence to a specific protocol. Detailed information regarding the Columbus protocol, last updated on January 27, 2021, can be accessed at https://data-onderwijs.vlaanderen.be/documenten/bestand.ashx?id=13051.

Data, Materials and Code availability

The datasets analyzed during the current study are not publicly available due to privacy concerns of sensitive data, but are available from the corresponding author on reasonable request. We can provide information to the editor or other researchers as to how the data was obtained, which variables were used, and any selection criteria for inclusion in the sample.

We will provide data documentation (such as coding) that was used to produce the results of this study upon request with other researchers.

Authors' contribution statements

All authors contributed to the study conception and design. Items were developed by Elisabeth Roels. Data was collected by a longitudinal project of the Flemish government. Data preparation and analysis were performed by Sofie Van Cauwenberghe. Methodology was elaborated by Sofie Van Cauwenberghe, Stijn Schelfhout and Nicolas Dirix. The first draft of the manuscript was written by Sofie Van Cauwenberghe and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

In a higher education system characterized by open access and a growing number of enrollments, it is essential to adequately inform students about their skills and interests to facilitate informed decision-making regarding their academic pursuits and minimize the risk of dropout (Fonteyne, 2017). However, the prevalence of dropout underscores the need for intervention strategies, prompting the provision of exploratory tools to secondary school students in their final year. Study orientation tools try to focus on cognitive knowledge on one hand, and encompass non-cognitive skills and interests on the other hand, as both factors contribute to academic success in higher education (Richardson et al., 2012; Fonteyne et al., 2017; Author et al., 2022).

Amongst these factors, intelligence arguably stands out as the pivotal predictor of academic performance (Roth et al., 2015). As a key part of the study choice process, intelligence assessment can hold a crucial role in guiding students towards higher education. Unfortunately, most available intelligence tests (such as the Stanford-Binet and Wechsler scales) are costly, limiting their accessibility in research settings. Additionally, these tests are time-consuming to administer, which could result in fatigue and reduced participant engagement (Ackerman & Kanfer, 2009). While freely available public domain intelligence tests exist, such as the International Cognitive Ability Resource (ICAR16) and Cog15, these measures are still under-researched, and their psychometric properties require further validation (Kajonius, 2014; Kirkegaard & Nordbjerg, 2015, Kristjánsdóttir & Zaiter, 2023). Furthermore, the overuse of established tests like Raven's Progressive Matrices has raised concerns about compromised validity due to test exposure, particularly in online research settings (Benisz et al., 2018). Freely available alternatives can help mitigate these issues by diversifying the pool of reliable assessment tools. However, despite the availability of some public domain measures, the need remains for additional tests that are both psychometrically sound and easily accessible. To address these challenges, we developed a free, short non-

verbal reasoning test designed to serve as an alternative measure of cognitive ability. By offering an additional public domain tool, we aim to expand the available resources for researchers, improve accessibility, and contribute to the ongoing refinement of intelligence assessment in psychological research. Key considerations in this test development include (1) the need to avoid cultural and verbal influences by using non-verbal items, (2) ensuring a limited testing time, (3) mitigating gender differences, (4) aligning the test structure with existing validated assessments and (5) free availability without fees that may hinder access for vulnerable students or countries. Rules is designed to identify students at risk for academic achievement, rather than to select top-students, and is therefore by design more sensitive at the lower end of the ability spectrum. From an orientation perspective, the goal is to assess basic reasoning skills and identify individuals who clearly lack these core cognitive abilities.

As the primary research goal, the present study evaluates the psychometric properties of the non-verbal reasoning test that we called Rules. We analyzed the test structure using confirmatory multidimensional item response theory (MIRT) models. To assess construct validity, participants completed both Rules and Raven's 2 Progressive Matrices Short Screener. Additionally, scores from a standardized mathematics and language proficiency test were collected to examine predictive validity. We do not claim to provide a comprehensive understanding of intelligence (or '*g*') through this test, but rather aim to efficiently assess a significant component of cognitive ability.

Cognitive Ability

Gignac (2018, p. 440) operationally defines human intelligence as “a person’s maximal capacity to complete a novel standardized task with veridical scoring using perceptual-cognitive processes.”

The concept of cognitive ability is extensively studied across various disciplines. Within the field of psychometrics, it is particularly utilized as a key metric to assess and

quantify this ability in relation to individual differences. Cognitive ability serves as a metric for one's adeptness in learning or processing information (Gottfredson, 1997). Learning is inherently shaped by intelligence, whether derived from experiences or other sources (Sternberg & Kaufman, 2011). Individuals with higher cognitive abilities exhibit a propensity for accelerated learning, irrespective of the source of information. They demonstrate enhanced capacity to retain data in their working memory, thereby facilitating efficient acquisition of knowledge and skills (Derue et al., 2012). The concept of cognitive ability emerges as a crucial predictor of success in both training and professional environments, particularly in roles involving higher levels of complexity. Gottfredson (1997) presents evidence supporting the pervasive usefulness of cognitive ability, asserting that its essence lies in the capacity to manage cognitive complexity, particularly in handling intricate information processing tasks.

Cattell-Horn-Carroll (CHC) Theory of Intelligence

In the context of the present study, we focus on fluid intelligence in accordance with the theory of intelligence of Cattell-Horn-Carroll (CHC).

~~The CHC model has its roots in several older intelligence theories. Following a period in which intelligence was conceptualized as a single underlying construct, referred to as the g-factor (Spearman, 1904), Cattell (1941, 1957) introduced a two-factor model of intelligence: fluid intelligence (Gf) and crystallized intelligence (Gc).~~ Cognitive ability is commonly categorized into fluid and crystallized intelligence, with crystallized intelligence denoting the reservoir of accumulated knowledge. Conversely, fluid intelligence represents an individual's capacity to manipulate both existing and novel information. Fluid intelligence denotes the capacity to tackle unfamiliar problems, engage in abstract reasoning, and navigate novel situations. This cognitive faculty encompasses a spectrum of skills, including pattern recognition, abstract reasoning, and problem-solving. ~~Subsequently, in the 1980s and 1990s, Cattell's Gf and Gc model was further expanded to include factors such as visual perception,~~

~~short-term memory, long-term storage and retrieval, processing speed, auditory processing, quantitative skills, and reading and writing skills (Horn, 1988, 1991; Horn & Noll, 1997).~~ After an extensive analysis, Carroll (1993, 1997) developed a hierarchical model of intelligence (the Three-Stratum Theory), marking the birth of the CHC model (Schneider & McGrew, 2012).

The CHC model consists of three levels. At the third level (Stratum III), the narrow, specific cognitive abilities are located. These include skills such as listening ability, general knowledge, and perceptual speed (Carroll, 1993; Schneider & McGrew, 2012). Stratum II logically represents a step higher in the hierarchy, encompassing eight to ten broad cognitive abilities. Alongside fluid and crystallized intelligence, other abilities such as short-term memory, processing speed, and auditory processing are grouped under Stratum II (Schneider & Newman, 2015). At the top of the hierarchy, in Stratum I, lies the overarching factor g, which describes general cognitive ability (Carroll, 1993, 1997; McGrew, 2009; Schneider & McGrew, 2012).

~~The positive manifold refers to the consistent pattern of positive correlations found among diverse cognitive ability tests (Kovacs & Conway, 2019). Originally identified by Spearman (1904), this phenomenon highlights the interrelated nature of various cognitive tasks. Essentially, all subtests that measure different dimensions of cognitive functioning tend to show significant intercorrelations (Burgoyne et al., 2022), which collectively contribute to the formation of a general intelligence factor, often understood as a formative construct in psychometric theory.~~

Within the CHC model, reasoning ability stands out as an important and specific cognitive skill as part of fluid intelligence. Reasoning ability entails the capacity to engage in logical thinking, comprehend abstract concepts, and employ problem-solving strategies effectively (Kaufman et al., 2016).

While the CHC model, which differentiates between fluid and crystallized intelligence, remains the prevailing framework in cognitive ability research, alternative models have been proposed. Vernon (1965) introduced a two-factor structure consisting of verbal/educational (v:ed) and spatial/mechanical (k:m) abilities. In contrast, Johnson and Bouchard (2005) suggested a three-factor model that includes verbal, perceptual, and image rotation abilities. Another perspective argues that these three components—verbal, perceptual, and rotation—are the most accurate way to describe cognitive abilities at a level below general intelligence (g) (Bouchard, 2014; Johnson & Bouchard, 2005).

Fluid Intelligence and Academic Achievement

The meta-analysis of Roth et al. (2015) demonstrated a population correlation of r = .54 between cognitive ability and school performance. CHC key studies highlight the importance of both broad and narrow cognitive abilities in predicting academic achievement, even when controlling for the overarching factor of general *g* (Flanagan, 2000; McGrew et al., 1997). Authors like Flanagan (2000) and also McGrew et al. (1997) suggest that g's impact on achievement is best comprehended as an indirect effect, mediated by various broad and narrow cognitive abilities.

Fluid intelligence is typically gauged through performance tests like the well-known (but not free) Raven's test (Raven et al., 1998), and emerges as a pivotal determinant of academic achievement (Kuncel et al., 2004), with studies consistently revealing its positive correlation with academic grades (Colom & Flores-Mendoza, 2007). The effect is particularly striking in its association with mathematics grades (McGrew & Wendling, 2010; Peng et al., 2019; Primi et al., 2010). More specifically, study outcomes suggest that certain abilities such as fluid intelligence play a pivotal role in the development of specific reading and math skills. The meta-analysis of Peng and colleagues (2019) found that Gf was moderately related to reading, r = .38, and mathematics, r = .41. Fluid intelligence emerge as crucial factor above

and beyond the predictive power of general cognitive and achievement constructs. If fluid reasoning skills are indicative of future performance and can be assessed through relatively brief tests, then it is certainly advantageous to incorporate these tests within the framework of academic study orientation. Therefore, the present study investigates this relationship.

Fluid Intelligence and Sex

According to Barel and Tzischinsky (2018), men excel in their ability to handle visual-spatial images in working memory. Women have an advantage in tasks requiring the retrieval of information from long-term memory and the acquisition and use of verbal information. Conversely, the study by Colom and Garcia-Lopez (2002) indicates that while males outperform females on the Raven test, there is no systematic difference favoring either sex in measures of fluid intelligence (Gf). Abad et al. (2004) confirmed the potential bias in the Raven's test, suggesting that males may perform better on certain items due to the fact that the items for reasoning are heavily relying on visuo-spatial processing.

Further research demonstrated mixed results, with some results in support of the notion that inductive reasoning tests with figural material tend to favor males, (e.g., Luo et al., 2021), while others (Piraksa et al., 2014; Waschl & Burns, 2020) found no significant differences. However, a recent meta-analysis by Waschl and Burns (2020) suggests that although effect sizes are generally small, males tend to have an advantage on inductive reasoning tests that involve figural material.

To address this issue, we aim to reduce sex differences within our test by minimizing the visuospatial complexity of reasoning items. Specifically, we modify problem structures to focus more on abstract relational reasoning rather than spatial transformations, which are known to contribute to male advantages in traditional inductive reasoning tasks. However, since Rules includes non-verbal items involving induction and deduction, a small advantage

for men may still persist. We strive to minimize this as much as possible while maintaining the integrity of the test's measurement of fluid intelligence.

To address this concern, we aim to minimize sex differences in our test by reducing the visuospatial complexity of reasoning items. Instead of heavily relying on spatial transformations, our approach emphasizes abstract relational reasoning, which does not inherently favor one sex over the other. While Rules still includes non-verbal items involving both inductive and deductive components, we have designed the items to limit undue advantage based on visuospatial processing skills. Nonetheless, some small performance differences may persist, but our goal is to mitigate these as much as possible while preserving the validity of the test as a measure of fluid intelligence.

Fluid Intelligence and SES

Research of Rindermann and colleagues (2010) discusses the influence of parental socioeconomic status (SES) and education on intelligence. While the correlation between SES, education and crystallized intelligence (Gc) tends to be stronger, fluid intelligence (Gf) is also associated with SES, although to a lesser extent (Anum, 2022). This finding means that the knowledge and skills acquired through education and life experience (crystallized intelligence) are more influenced by parents' socioeconomic and educational background (Anum, 2022). Fluid intelligence is more biologically based and less influenced by environmental factors. However, the findings also support the idea that fluid intelligence is not entirely immune to environmental influences, and conversely, higher levels of fluid intelligence may also shape the environment, as individuals with higher intelligence may actively seek out enriching experiences and opportunities (Trapp & Ziegler, 2019). It is important to note that this association is bidirectional: while SES may influence intelligence, it is also possible that intelligence contributes to higher SES through pathways such as educational attainment and career choices (Duyck, 2023). This bidirectional understanding

helps to avoid the assumption that intelligence tests are unfairly biased by SES. In the present study, we examine SES in relation to Gf and academic performance in reading and mathematics, acknowledging that intelligence and SES are correlated, with typical medium to strong correlations around $r = .38$ (Levine, 2011). Additionally, research highlights the importance of distinguishing SES from other factors, such as maternal cognitive abilities, when making such claims (Marks & O'Connell, 2021a, 2021b).

Public Domain Testing

Intelligence tests are widely utilized to assess an individual's mental capacities and compare them to those of others using scaled scores (Braaten & Norman, 2006). These tests rank among the most accurate psychological measurement instruments, still renowned for their reliability and validity and highly valued by experts (Gottfredson, 1997; Rindermann et al., 2020; Bloemink, 2023). Meta-analytic test–retest reliabilities of the test scores varied from adequate to high, with correlations of $r = .70$ and above (Calamia et al., 2013). The outcome of an intelligence test is typically expressed as an IQ score, which relates an individual's performance to others within the same age group. IQ scores are presumed to follow a normal distribution within the population.

The advantages of utilizing free public domain resources for researchers are described by Condon and Revelle (2014). These benefits include cost-effectiveness, increased control over test content, and the potential for a more nuanced understanding of the correlation structure between constructs. Additionally, public domain measures offer a collaborative platform for researchers to contribute to test development, refinement, and validation, ultimately benefiting the research community by facilitating empirical comparisons across diverse criteria. While public domain alternatives exist, including the International Cognitive Ability Resource (ICAR16) and Cog15, they each have limitations that warrant the development of additional assessments. The ICAR16, a freely available cognitive ability

measure, has demonstrated good psychometric properties, particularly in assessing fluid reasoning and visual-spatial processing (Young & Keith, 2020). However, studies indicate that the instrument remains under-researched, with concerns about its generalizability across diverse populations (Kirkegaard & Nordbjerg, 2015). Similarly, the Cog15 is a brief cognitive ability test that has yet to be validated (Kajonius, 2014), raising questions about its reliability and construct validity. Moreover, the overuse of established tests like Raven's Progressive Matrices has raised concerns about compromised validity due to test exposure. Freely available alternatives can help mitigate these issues by diversifying the pool of reliable assessment tools. To address these challenges, we constructed a new, freely available, and concise non-verbal reasoning test. Our goal is to provide researchers with an alternative measure that enhances accessibility while maintaining strong psychometric properties. This test is designed to serve as both a standalone assessment of non-verbal intelligence as well as a control variable in various research contexts, further expanding the range of reliable public domain intelligence measures available to the scientific community.

Present Study

The present paper features three studies where we evaluate Rules, a free measure for non-verbal reasoning. Note that Rules does not measure IQ as most of the CHC theory's broad abilities are not covered by the scale. In other words, Rules estimates the fluid (Gf) aspects of cognitive ability, rather than the broad CHC concept of IQ.

The first study evaluated (1) the internal consistency of the subtests and Rules as a whole, (2) the distribution of Rules scores and the distribution over sex and SES by differential item functioning (DIF), (3) item characteristics explored by Item Response Theory, and (4) structural properties of a 28-item Rules measure. We used multidimensional item response theory for confirmatory models (correlated dimensions model and bifactor model) by using maximum-likelihood measures. We examine the structural validity of the

non-verbal reasoning test Rules, through the framework of the most well-supported theory of intelligence, Cattell– Horn–Carroll (CHC) theory (McGrew, 2009).

The second study evaluates the construct validity of Rules' items when administered online, by cross-validating with a brief commercial measure of cognitive ability, Raven's 2 Progressive Matrices Short Screener (McLeod & McCrimmon, 2021). The relationship between Raven's 2 scores and performance on comparable cognitive functioning measures demonstrated moderate to strong correlations (McLeod & McCrimmon, 2021). A literature review revealed a weighted average correlation of .67 between the WAIS and Raven (*Correlation Between The Wechsler Adult Intelligence Scale And Raven's Progressive Matrices*, 2018; McLeod & Rubin, 1962).

The third study evaluated the predictive validity of Rules for academic performance by examining the relation with scores on a standardized mathematical skills and language test in the context of a large-scale online self-assessment tool that students in secondary education take for orientation towards higher education (the nationwide Columbus tool). We hypothesize that fluid intelligence (Rules) will serve as a significant predictor of both mathematical skills and language proficiency. Based on the findings of Peng et al. (2019), which identified moderate correlations between Gf and reading (r = .38) as well as Gf and mathematics (r = .41), we expect that higher scores on Rules' tasks associated with Gf will positively correlate with higher scores on standardized tests of mathematics and language skills.

Method

Data and Procedure

Data were gathered from two samples: a secondary database of N_1 = 32,585 students (60% female) in their last year of secondary school gathered within the longitudinal Columbus project (period 2016-2020) for study 1 and 3, and a primary sample of N_2 = 235

(47% female) last-year students of general (44%), technical (29%) and vocational (27%) secondary education who also completed the non-verbal tests Rules and Raven's 2 Short Screener (between February and May 2023), in a counterbalanced order for study 2.

Secondary data were collected from Columbus: a long-term study orientation initiative initiated by the government for prospective students in Flemish higher education (Demulder et al., 2020). Columbus is a large scale online self-assessment and feedback instrument, attempting to improve study orientation between secondary education (especially general and technical education[1]) and higher education. Although we used a convenience sample, participants were primarily recruited within school settings, often as part of broader educational guidance procedure. As a result, participation was largely institutionally driven rather than self-initiated. As a non-verbal reasoning test, Rules is part of this cognitive test battery, in addition to a mathematical skills test and an academic language proficiency test. For the present paper, data were used from students' reasoning, math and language ability in the last year of secondary education extracted from cohorts 2016-2017, 2017-2018, 2018-2019 and 2019-2020. The participants in the study had an average age of 18 years old, with ages ranging from 17 to 21 years. The distribution of participants across SES categories was as follows: 61.3% were classified as SES = 0 (high SES), 24.7% as SES = 1, 8.4% as SES = 2, 4.0% as SES = 3, and 1.6% as SES = 4 (low SES).

Compliance with Ethical Standards

The present study is part of "The Columbus Project: a tool for study orientation towards higher education", to which the Ethical Commission of the Faculty of Psychology and Educational Sciences at Ghent University has given approval with reference 2016/82.

[1] Columbus data consist of 61.4% general, 36.1% technical, 1.3% vocational and 1.3% artistic secondary education students.

Measures

Raven's 2 Progressive Matrices Short Screener

The *Raven's 2 Progressive Matrices Screener* serves as a non-verbal intelligence test designed to rapidly assess general *g* (Pearson, 2020; Raven & Raven, 2018). Based on data from the United States, the marginal reliability coefficient for the Raven's 2 Digital Short Form is reported to be .80 (Dimitrov, 2003; McLeod & McCrimmon, 2021). The test assesses deductive reasoning, a key component of general *g* as identified by Spearman (1904). Deductive reasoning encompasses the ability to derive new insights, extract meaning from complexity, identify patterns, and establish connections (Raven et al., 2018; Pearson, 2020).

Participants are required to discern a rule based on provided information and apply this rule to the missing section. When administered digitally through Q-Global, the Raven's 2 Progressive Matrices Short Screener consists of 24 items, which are drawn from a larger item bank of 329 items. These 24 items are selected to provide a representative measure of deductive reasoning and are presented with straightforward and concise instructions. The test requires approximately 20 minutes per participant. A score report was obtained via Q-Global (Pearson, 2023). The report describes a scaled score which is a standardized score, and can be compared to other intelligence tests. Administering Raven is not free, 25 pre-paid digital test administrations cost €104.19 (Pearson, 2023).

Featuring a wide age range from 4 to 69 years and covering a broad spectrum of cognitive abilities (IQ 40-160), the Raven's test is applicable to diverse populations, including individuals with communication disorders, limited verbal capacities, non-native speakers of Dutch, or those who are deaf or hard of hearing.

Rules

We developed a Raven-style non-verbal reasoning test, *Rules,* to operationalize fluid intelligence. The ~~reliability and~~ validity of a measurement depends on both the instrument and

the sample used. Therefore, we assessed the internal consistency of all measures for the current sample (Harris, 2003; Graham, 2015).

As deduction and induction are generally considered as indicators of Gf (McGrew, 2009), the test consists of 28 items in MC-format, with 14 items about deductive and inductive reasoning each. The first segment, termed 'deduction', presents participants with a series of drawings, wherein one deviates from the others (see example Figure 1). Participants are tasked with identifying this divergent drawing by discerning its distinguishing characteristic and implying rules. In contrast, the second segment of Rules, labeled 'induction', presents participants with a prototype drawing that undergoes a transformation to generate a subsequent drawing by applying the same rules (see example Figure 2). Participants are subsequently provided with a new drawing and are required to replicate the transformation by analogy with the prototype. Selection of the appropriate alteration is made from a set of four possible options. The 28-item Rules can be found in Appendix A.

Figure 1

Example of Deduction Item

Figure 2

Example of Induction Item

Rules is designed as a research and counseling assessment instrument in an open access study environment with low tuition fees and no performance prerequisites like obtaining minimal exam scores or grades. Rules aims to be discriminative at the lower end of the ability spectrum of higher education. From an orientation perspective, the primary goal is to assess basic skills and identify individuals who clearly lack these core cognitive abilities for successful achievement in higher education (International Standard Classification of Education (ISCED) levels 6 and 7). The test is not intended to select or predict performance at the higher end of the spectrum.

Cognitive Measures

As mathematical competence is one of the most important factors for academic achievement, we assess basic mathematical skills (Fonteyne et al., 2014) using 25 items. Advanced numerical competence is measured by 10 additional questions as some academic degrees require higher levels of mathematical insight. The basic and advanced mathematical test scores can be combined into one *mathematics score* (α = .89) with items both in MC-format and open questions. One example question is "In a group of 400 people, there are 270 men and 130 women. Among the women, the proportion of low-skilled individuals is 0.4. How many low-skilled women are there specifically?".

Academic Language Proficiency is tested with the Short Academic Reading and Vocabulary test (SARV) (Heeren et al., 2020). The 14 items (α = .66) test word knowledge and reading ability. Word knowledge items test whether word meaning can be inferred from the context or word form. Reading ability tests insight into text construction where students must recognize the major text structures and be able to pick the correct one-sentence resume of a text in a multiple-choice format. There is a time limit to encourage students' strategies (Hulstijn, 2015).

Sex

Sex (male/female) was included as a binary variable (0/1) obtained by linking the data to the administrative database of the Flemish Department of Education and Training, which contains information about students' secondary education careers.

Socio-Economic Status

Socio-economic status (SES) in an educational context was measured by the educational disadvantage indicator which represents the student's social profile (Flemish Government, 2018; Avvisati, 2020) for the Flemish government to determine access to (financial) support measures. A number ranging from zero to four is computed by indicating

whether 1) the mother's educational level is not higher than secondary education (11.6% of the sample received a point on this indicator), 2) the language spoken at home is different from the language spoken at school (8.5% of the sample), 3) the neighbourhood has a high percentage of 15-year-olds with a school delay of two years or more (16.1% of this sample) and 4) the student receives a scholarship (23.8% of this sample). Higher scores indicate greater disadvantage.

Analyses

Analyses were conducted in SPSS (IBM SPSS Statistics, Version 29) and RStudio (version 1.3.1093). The internal consistency of the various subtests, as well as the overall Rules, is evaluated using Cronbach's alpha and McDonald's omega (Dunn et al., 2014; McDonald, 1999). To determine the internal consistency, we follow the COTAN framework for test quality assessment (Evers et al., 2010). This framework provides specific threshold criteria for evaluating internal consistency, which differ based on the test's intended application. Our primary objective is to develop a tool for research purposes, we apply the cut-off values established for group-level studies. Under these guidelines, internal consistency coefficients above .70 are considered good, those between .60 and .70 are acceptable, and coefficients below .60 are regarded as inadequate (Nunnally & Bernstein, 1994, p. 265).

The correlations between Rules and the subtests were adjusted to account for overlap due to shared error variance (Bashaw & Anderson, 1967; Cureton, 1966). This process, executed using the scoreOverlap function, removes the variance contributed by overlapping items and replaces it with the most accurate estimate of common variance, which is the squared multiple correlation for each item (Revelle, 2024).

A *t*-test and anova were conducted to investigate the differences for sex and socio-economic status (SES) on test level. To ensure that the test items did not exhibit bias towards students with specific background characteristics, a Differential Item Functioning (DIF)

analysis was conducted for sex and SES. The DIF analysis followed the guidelines proposed by Strobl et al. (2015), investigating potential sources of DIF, such as item wording, characteristics, or multidimensionality in the test.

To assess both the significance and magnitude of DIF, the Mantel-Haenszel Chi-square test was employed, and deltaMH was calculated with the difR package (Magis et al., 2020). These procedures are based on the Mantel-Haenszel[2] DIF method as described by Magis et al. (2010, 2020). The structural validity of the non-verbal test was assessed using confirmatory Multidimensional Item Response Theory (MIRT) conducted by the "mirt" package in R (Chalmers, 2012). MIRT is an extension of IRT and is used to explore and validate underlying test structure and dimensionality (Immekus et al., 2019; Kruglova & Dykhovychnyi, 2022). When the test assesses more than one underlying ability, MIRT models such as confirmatory (Embretson & Reise, 2000) are employed to investigate and assess whether they adequately represent the structure of the CHC framework. Analyses based on three-parameter MIRT were used to evaluate multidimensional relationships between items on several levels, including (1) all 28 items, (2) the two item types – deduction and induction – independently, and (3) the items with a second order g-factor. Pseudo-guessing parameters are freely estimated. The fit of the model was assessed using various fit statistics[3], including the chi-square statistic, RMSEA, comparative fit index (CFI), and standardized root mean square residual (SRMR) to select the best fitting model. In this paper, we compared two MIRT models: one with two correlated dimensions and a second-order factor model.

[2] The absolute value of deltaMH serves as an effect size indicator for DIF, with classification following the ETS criteria: a negligible effect when $|\Delta MH| \leq 1$ (Class A), a moderate effect when $1 < |\Delta MH| \leq 1.5$ (Class B), and a large effect when $|\Delta MH| > 1.5$ (Class C) (Magis et al., 2020).

[3] Hu and Bentler's (1999) derived cutoffs for evaluating model-data fit using Maximum Likelihood (ML) estimation and provide specific guidelines for interpreting key fit indices: RMSEA values below 0.05 indicated a good fit, while values around 0.08 suggested reasonable fit. CFI values exceeding 0.95 and SRMR values below 0.08 were considered indicative of acceptable fit.

Construct validity was evaluated through the bivariate correlation between our non-verbal test and the Raven's 2 Progressive Matrices short screener, a well-established measure of fluid intelligence (Raven & Raven, 2018). The correlation analysis provided insights into the extent to which the non-verbal test aligns with another measure of intelligence, thereby assessing its construct validity.

Predictive validity was assessed through linear regression analyses conducted in IBM SPSS Statistics (version 29). These analyses aimed to investigate the extent to which scores on the non-verbal Rules test predict academic performance measured by mathematics scores and academic language proficiency. This analysis enabled us to determine whether the non-verbal test scores could effectively predict academic performance.

Results

Study 1: Structural Validity

Internal Consistency

The internal consistency was measured for the subtests, deduction and induction, and for Rules as a whole by calculating Cronbach's alpha (Reise & Haviland, 2024; Revelle & Condon, 2019; Taber, 2017) and McDonald's omega[4] (Dunn et al., 2014; McDonald, 1999). The overall Rules test showed good internal consistency with a Cronbach's $\alpha = .79$ and McDonald's $\omega = .80$. The items of the deduction subtest showed a Cronbach's alpha of $\alpha = .63$ and McDonald's omega of $\omega = .65$, which was acceptable (Evers et al., 2010). The internal consistency for the subtest induction was good with a Cronbach's alpha of $\alpha = .70$ and McDonald's omega of $\omega = .72$. By examining the correlation between the two subtests, we determine that Revelle's β is .75, which serves as an estimate of the general factor saturation of the test. This metric provides a more informative representation of the test's

[4] When the assumption of tau-equivalence is not met—a common occurrence in psychology—omega proves to be more reliable than alpha (Dunn et al., 2014; Zinbarg et al., 2005). Since omega reduces the likelihood of either overestimating or underestimating reliability, it is considered the superior option.

internal structure (Zinbarg et al., 2005). 75% of the common variance can be attributed to a single latent trait.

Distribution of Rules Scores

The distribution of Rules scores is shown in Figure 3. We converted the original total scores ($M = 20.11$, $SD = 4.51$) into z-scores ($M = 0$, $SD = 1$), revealing a distribution that displayed a slight leftward skew, which confirms the intended focus on sensitivity towards the lower end of the ability scale.

Figure 3

Distribution of Standardized Rules z-scores

To investigate any sex differences, we conducted an independent samples *t*-test with sex as grouping variable and total score of Rules as dependent variable. We found a small significant difference between boys ($M = 20.98$, $SD = 4.49$) and girls ($M = 19.54$, $SD = 4.43$); $t(32643) = 28.47$, $p < .001$, Cohen's $d = 0.32$.

At item level, we conducted DIF analyses. The deduction part of the Rules test contains 10 items that exhibit significant sex-related DIF ($p < .01$). Of these, nine are classified as A-items with a negligible effect, and one as a B-item (ded12) with a moderate effect. The B-item appears to be easier for girls (see Appendix B, Table B.1, Figure B.1). The induction part of the Rules test contains 11 items that exhibit significant sex-related DIF ($p < .01$). All items are classified as A-items with a negligible effect (see Appendix B, Table B.1, Figure B.2).

We wanted to test if there is a difference on reasoning results depending on the SES of students. The linear regression analysis showed that SES had a significant effect on Rules, $\beta = -0.89$, $t(32622) = -33.326$, $p < .001$. The model explained 3% of the variance in Rules, adjusted $R^2 = .03$. There was a statistically significant difference between SES categories as determined by one-way ANOVA ($F(4,32619) = 280.25$, $p < .001$, $\eta^2 = .03$). Games Howell

post hoc tests revealed that Rules scores for students with high SES are significantly higher compared to students with low SES ($p < .001$). On average, students with high SES score 0.74 points higher on Rules than students who mark one of the four SES variables (education mother, language, neighbourhood, scholarship), 1.77 points higher than students who mark two, 2.85 points higher than those who mark three, and 3.55 points higher than students who mark all SES variables (low SES).

At item level, we conducted DIF analyses. Four deduction items show significant DIF related to SES ($p < .01$). All of these items were classified as A-items, indicating negligible effects (see Appendix B, Table B.2, Figure B.3). Two induction items show significant DIF related to SES ($p < .01$), but were classified as A-items (see Appendix B, Table B.2, Figure B.4).

Adherence to CHC Model

In Table 1, the correlations have been adjusted for reliability utilizing the standardized alpha specific to each scale (Bashaw & Anderson, 1967; Cureton, 1966; Revelle, 2024). Correlations without correcting for scale reliability can be found in Table 2.

Table 1

Correlations of Rules and Subtests Corrected for Item Overlap and Attenuation

	1	2	3
1. Deduction	.63	.89	.97
2. Induction	.59	.71	.97
3. Full Rules test	.69	.73	.79

Note. Corrected correlations below the diagonal, alpha on the diagonal, corrected correlation for attenuation above the diagonal. $N = 32{,}585$, $p < .01$.

Table 2

Raw Correlations of Rules and Subtests Corrected for Attenuation

	1	2	3
4. Deduction	.63	.90	1.26
5. Induction	.60	.70	1.21
6. Full Rules test	.89	.90	.79

Note. Raw correlations below the diagonal, alpha on the diagonal, corrected correlation for attenuation above the diagonal. $N = 32{,}585$, $p < .01$.

Item Characteristics

IRT analysis provided a test information curve for the 28 item Rules test. In Figure 4 is shown that especially the skills of those who score lower are well mapped out: 67.3% of the test information lies between -4 and 0. Difficulties ranged from -3.27 to 2.12, with a mean difficulty of -1.00 (see Table 3). Figure 5 shows the conditional reliability of Rules to indicate how precisely the test measures subgroups of persons or at various cut scores. Rules has more measurement errors within the subgroup characterized by high latent ability.

Figure 4

Test Information Function for the 28 Item Rules Non-Verbal Test

Note. 67.3% of the test information lies between -4 and 0. $N = 32{,}585$.

Figure 5

Conditional Reliability for the 28 Item Rules Non-Verbal Test

Note. Conditional reliability coefficients indicate the reliability for subgroups at various cut scores on a test. Rules shows more measurement errors in the high latent ability subgroup. $N = 32{,}585$.

Confirmatory Factor Analysis

We conducted a confirmatory factor analysis using a multidimensional item response theory (MIRT) model to explore the underlying structure of a 28-item reasoning test. Item responses were scored dichotomously as correct (1) or incorrect (0). We hypothesized that (1) the test measures two distinct dimensions of fluid reasoning ability: deductive reasoning and inductive reasoning, and (2) that these dimensions load on the higher-order g fluid.

The MIRT model with correlated dimensions ($\chi^2 = 2618.49$, df = 28) demonstrated adequate fit to the data, with a Comparative Fit Index (CFI) of 0.99, Root Mean Square Error of Approximation (RMSEA) of 0.02, and Standardized Root Mean Square Residual (SRMSR) of 0.02, indicating marginally above the cutoff for acceptable model fit. The correlation between the deduction and induction item types was r = .89, 95% BI [.876, .895].

The MIRT model with correlated dimensions which loaded on a higher-order dimension ($\chi^2 = 2627.58$, df = 28) demonstrated good fit to the data, with a Comparative Fit Index (CFI) of 0.99, Root Mean Square Error of Approximation (RMSEA) of 0.02, and Standardized Root Mean Square Residual (SRMSR) of 0.02, indicating a good model fit. Thus, the higher-order model provides the best fit to the data. Coefficients of the second order MIRT model are described in Table 3. All factor loadings can be found in Appendix B Table 3 and 4.

Table 3

Coefficients of Second Order MIRT Model

	a1	a2	a3	d	b	g	u
Ded1	0.74	0.74	0	1.39	-1.88	0.00	1
Ded2	0.76	0.76	0	0.46	-0.61	0.10	1
Ded3	0.80	0.80	0	0.29	-0.36	0.00	1
Ded4	0.85	0.85	0	1.19	-1.40	0.00	1

Ded5	0.96	0.96	0	3.01	-3.14	0.00	1
Ded6	0.98	0.98	0	-2.08	2.12	0.06	1
Ded7	1.02	1.02	0	-1.31	1.28	0.24	1
Ded8	1.11	1.11	0	1.50	-1.35	0.07	1
Ded9	1.11	1.11	0	0.13	-0.12	0.30	1
Ded10	1.16	1.16	0	3.79	-3.27	0.05	1
Ded11	1.23	1.23	0	3.02	-2.46	0.00	1
Ded12	1.30	1.30	0	-0.59	0.45	0.37	1
Ded13	1.32	1.32	0	0.34	-0.26	0.15	1
Ded14	1.34	1.34	0	0.95	-0.71	0.14	1
Ind1	0.76	0	0.76	2.10	-2.76	0.00	1
Ind2	0.96	0	0.96	1.57	-1.64	0.01	1
Ind3	1.01	0	1.01	1.95	-1.93	0.00	1
Ind4	1.13	0	1.13	0.30	-0.27	0.13	1
Ind5	1.14	0	1.14	0.49	-0.43	0.32	1
Ind6	1.14	0	1.14	-0.59	0.52	0.12	1
Ind7	1.14	0	1.14	1.88	-1.65	0.00	1
Ind8	1.19	0	1.19	2.92	-2.45	0.00	1
Ind9	1.22	0	1.22	2.60	-2.13	0.13	1
Ind10	1.23	0	1.23	2.78	-2.26	0.00	1
Ind11	1.30	0	1.30	-0.71	0.55	0.32	1
Ind12	1.51	0	1.51	1.88	-1.25	0.00	1
Ind13	1.53	0	1.53	1.33	-0.87	0.14	1
Ind14	1.76	0	1.76	-0.26	0.15	0.25	1

Note. Items starting with Ded are the deduction items, with Ind are the induction items. a1 = discrimination parameter for *g*; a2 = discrimination parameter for deduction; a3 = discrimination parameter for induction; d = intercept; b = difficulty parameter; g = guessing parameter; u = upper asymptote parameter. *N* = 32,585.

Study 2: Construct Validity

To investigate the construct validity, we conducted a bivariate Pearson correlation between the total score on Rules and the scaled score on Raven's 2 Progressive Matrices Short Screener (see Table 4). The uncorrected correlation between the respective overall observed scores, Rules and Raven total score, was moderate in magnitude ($r = .62$, $p < .01$). When corrected for attenuation due to measurement error, the estimated correlation increased to $r = .78$.

Table 4

Zero-Order Correlations between Rules and RPM

	1	2	3	4
1. RPM (scaled)	1			
2. Rules: Deduction	.54**	1		
3. Rules: Induction	.57**	.59**	1	
4. Full Rules test	.62**	.88**	.91**	1

Note. $N = 235$, ** $p < .01$

Study 3: Predictive Validity

We computed Pearson correlations between the scores obtained from Rules (sub)tests and various indicators of academic performance (see Table 5). The results indicated significant positive correlations between all measures of academic performance and both the individual subtests as well as the complete Rules test.

We conducted two hierarchical linear regression analyses to investigate the impact of Rules total score on a standardized mathematics test and an academic language proficiency test after controlling for socio-economic status (SES) and sex. Both models were statistically significant and accounted for a substantial proportion of the variance in the dependent variables. The results showed that the model with SES and sex was significant for

mathematics ($F(2, 31319) = 1664.60$, $p < .001$, adj. $R^2 = .10$) and for language ($F(2, 30859) = 549.81$, $p < .001$, adj. $R^2 = .03$). Both SES and sex were significantly associated with scores on a standardized mathematics and language test.

For mathematics, the model ($F(3, 31318) = 6946.41$, $p < .001$, adj. $R^2 = .40$), which included total Rules score ($b = 0.57$, $t = 125.808$, $p < .001$) showed significant improvement from the model with only SES and sex $\Delta F(1, 31318) = 15827.67$, $p < .001$, $\Delta R^2 = .30$.

For language, the model ($F(3, 30858) = 2237.11$, $p < .001$, adj. $R^2 = .18$), which included total Rules score ($b = 0.39$, $t = 73.612$, $p < .001$) showed significant improvement from the first model $\Delta F(1, 30858) = 5418.67$, $p < .001$, $\Delta R^2 = .14$. The detailed results of the full regression analyses are presented in Table 6.

Table 5

Zero-Order Correlations between Rules (Sub)Tests and Academic Performance

Academic Performance	Deduction	Induction	Full Rules
Mathematics	.53**	.55**	.61**
Language	.35**	.39**	.41**

Note. $N = 32{,}585$, ** $p < .01$

Table 6

Two Hierarchical Linear Regressions

Academic Performance (dependent)	Predictor	B	*t*	*p*	Unique R^2
Mathematics	Sex (0 = male)	-0.17	-37.415	<.001	
	SES (0 = high SES)	-0.07	-16.663	<.001	
	Rules total	0.57	125.808	<.001	0.37
Language	Sex (0 = male)	-0.04	-8.105	<.001	
	SES (0 = high SES)	-0.08	-15.965	<.001	
	Rules total	0.39	73.612	<.001	0.17

Note. Rules uniquely explains 37% of the variance in mathematics and 17% of the variance in language. Betas are standardized. N = 32,585.

Discussion

Cognitive ability assessment plays a pivotal role in the decision-making process for higher education pathways. However, many existing intelligence tests are either prohibitively expensive or require considerable time to administer. To address this issues, we developed Rules, a free, quick and easy to administer non-verbal fluid reasoning test. Although Rules is originally designed to pinpoint those who are highly likely to lack the skills needed to succeed in the first year of higher education, its applicability extends beyond this group. Rules can be effectively used in broader educational settings, offering a versatile tool for identifying reasoning abilities ~~across a wide range of cognitive levels~~ and as a control variable in population research where basic intelligence needs to be verified or ruled out. It is important to note that we do not intend for this test to provide a comprehensive measure of intelligence (or '*g*'), but rather to evaluate one crucial aspect of cognitive ability, just like the well-known Raven's test for instance.

The non-verbal reasoning scale Rules seems to be a promising public domain test to evaluate fluid intelligence and in particular deduction and induction. We conducted three studies: (1) we evaluated the internal consistency, the distribution and the structural validity of the test, (2) we examined the construct validity by cross-validating Rules with Raven's 2 Progressive Matrices, and (3) we investigated the predictive validity of Rules with a standardized mathematics and language proficiency test.

The two-subscale reasoning test displays adequate overall internal consistency with α =.79 and ω = .80, which is sufficient for population research (Evers et al., 2010). We determine that Revelle's β is .75, which serves as an estimate of the general factor saturation of the test. We recommend caution when using and interpreting the subtests independently, and always suggest testing reliability before use.

The distribution of scores on Rules showed a slightly leftward skew, likely due to range restriction on the high end of performance in the sample, a ceiling effect (see Cronbach, 1990, pp. 210-212). This skewness is by design, as the reasoning test was developed as part of a broader low-stakes study exploration tool aimed to identify students who may lack the necessary skills, but still are considering higher education. The primary goal of this tool was to provide information about entering higher education, rather than to identify individuals who excel in reasoning tasks. As such, the results are working as intended and are sufficient for population research, particularly in the context of educational orientation, but they are not suitable for selection purposes that require identification of the best performing individual.

We investigated sex differences on the non-verbal test Rules overall and we found a small advantage for boys with a small effect size of Cohen's d = 0.32. At item level, DIF analyses indicated that several items showed differential item functioning related to sex. However, the effects were negligible, except for one item (ded12), which displayed a moderate effect in favor of girls. In light of our findings that boys slightly outperform girls in

the non-verbal reasoning test, it is pertinent to note that this aligns with existing literature on gender differences in cognitive abilities (Colom & Garcia-Lopez, 2002; Abad et al., 2004). Barel and Tzischinsky (2018) highlight that men tend to excel in manipulating visual-spatial images in working memory, which is a crucial skill for tasks involving deduction and induction, the two subtests of our test. Furthermore, while Colom and Garcia-Lopez (2002) found no systematic difference in fluid intelligence between sexes, they did report that boys generally perform better on the Raven test, a well-known measure of non-verbal reasoning. These findings collectively support the notion that boys might have a slight advantage in our non-verbal reasoning test.

Regarding socio-economic status a very small effect of $\eta^2 = .03$ was found where students with high SES scored significantly higher compared to students with low SES. At item level, DIF analyses indicated that several items showed differential item functioning related to SES. However, the effects were negligible. Our findings align with Rindermann et al. (2010), who suggest that while parental SES and education have a stronger influence on crystallized intelligence, they still exert some influence on fluid intelligence. The slight advantage observed in high SES students on the non-verbal reasoning test can be attributed to these environmental influences. However, the small effect size in our study is encouraging for the test validity, indicating that the test is largely independent of socioeconomic factors and primarily measures inherent non-verbal reasoning ability.

Concerning the structural properties of our 28-item non-verbal measure of fluid reasoning, the correlation matrix and confirmatory Multidimensional Item Response Theory (MIRT) models indicated ~~the presence of a positive manifold in our data, demonstrating that~~ both subtests, deduction and induction, contribute distinctly to assessing fluid intelligence. These findings support the CHC theory. IRT analysis revealed that especially the skills of those who score lower are well mapped out: 67.3% of the test information lies between -4 and

0. We conducted a confirmatory factor analysis using a multidimensional item response theory (MIRT) model to explore the underlying structure of Rules. The MIRT model with correlated dimensions which loaded on a higher-order dimension demonstrated good fit to the data.

The second study evaluated the construct validity of Rules' items when administered online, by cross-validating Rules with Raven's 2 Progressive Matrices Short Screener in a sample of 235 last-year secondary school students. We found a bivariate correlation of .62, ~~further supporting the presence of a positive manifold.~~ The magnitude or extent of the relationship observed in the current study is in line with prior cross-validation research with Raven (McLeod & McCrimmon, 2021). However, it is important to note that Rules does not aspire to measure IQ in the same manner as the RPM scales. Instead, Rules aims to be a concise, public domain measure of cognitive ability aligning with the overarching goals of the Columbus project (Demulder et al., 2020) or as a control variable for research purposes.

Our third and last study focused on predictive validity of Rules. We found that fluid intelligence, as measured by Rules, could uniquely explain 37% of the variance in a standardized mathematics test and 17% in an academic language proficiency test. Additionally, it is important to note that we accounted for sex and SES, enhancing the robustness of our results. Moreover, the correlations observed in our study are comparable in magnitude to those reported by Peng and colleagues (2019) in their research on mathematics and reading.

This study offers several concrete contributions to the field of intelligence research. First, it introduces a freely accessible, nonverbal reasoning instrument, responding to the recognized need for open-access alternatives to proprietary tests such as Raven's Progressive Matrices. Second, the test development was guided by theoretical and empirical insights from the literature, including attention to gender-related item functioning—a psychometric concern

often neglected in legacy instruments. Third, by offering a brief and easily administrable tool, our study facilitates the integration of intelligence as a control or explanatory construct in large-scale, population-based research. In such contexts, access to an efficient and psychometrically sound reasoning test allows researchers to better isolate or confirm intelligence-related effects, thereby strengthening the construct validity of their findings. These contributions position our work not only within the practical realm of educational screening, but also within the theoretical and methodological core of intelligence research.

Limitations and Future Research

While Rules offers a promising, accessible tool for evaluating non-verbal fluid reasoning, several limitations must be acknowledged. A key limitation of Rules is the risk of item exposure through web searches, which could affect long-term validity. However, the abstract nature of the items reduces memorization risks, making exposure less problematic for applied use. Additionally, automatic item-generation techniques (Arendasy et al., 2006; Condon & Revelle, 2014; Dennis et al., 2002) could mitigate these concerns by creating large item pools with controlled difficulty. Future research should explore integrating such methods into public-domain assessments. Implementing ethical guidelines, such as warnings for nonprofessionals and providing sample items to discourage unauthorized distribution (Goldberg et al., 2006), can help maintain integrity. Further research should assess the actual impact of item exposure on validity and explore security measures suited to different testing environments.

A further consideration is the limited difficulty range of Rules. The test was designed to identify students who may struggle with the cognitive demands of higher education, not to assess the full range of non-verbal intelligence. Consistent with this goal, item response theory analyses showed that the test provides the most information for below-average scorers, with 67.3% of test information located between -4 and 0 on the latent ability scale. This

limitation is by design: Rules is not intended to assess the full range of cognitive ability or to function as a general-purpose intelligence test. Rather, it was developed as a screening tool to identify students who may lack the cognitive prerequisites for success in higher education, especially in an open access study environment, with low tuition fees and no performance requirements. As such, its utility lies in supporting educational transitions rather than high-stakes selection or identifying high-ability individuals. While this application narrows its scope, it provides practical value in informing targeted interventions and resource allocation. Future research should consider developing complementary items or test versions targeting higher ability levels, potentially expanding its utility for broader populations. However, reliability and validity depend not only on the instrument itself but also on the specific population being assessed. Therefore, further research is necessary to determine whether these subtests might yield more reliable results in different populations.

Finally, as Rules continues to be used in applied and research settings, future research should explore the longitudinal predictive validity of Rules, particularly regarding academic persistence and performance beyond the first year of higher education. Further validation efforts should also examine the test's applicability across different contexts to ensure broader generalizability. In parallel, continued psychometric development, such as adaptive testing formats or item bank expansion, may enhance the scalability and robustness of the instrument.

Despite these limitations, the current study offers a significant methodological contribution to the field of intelligence assessment by introducing and validating a concise, public-domain measure of non-verbal reasoning. By addressing accessibility, reliability, and practicality, Rules adds valuable diversity to the existing landscape of cognitive ability tests.

Conclusion

This research validated Rules, a free, short, and accessible non-verbal fluid reasoning test developed to identify students who may face challenges with the cognitive demands of

higher education. Across three studies, Rules demonstrated satisfactory internal consistency, strong structural and construct validity, and meaningful predictive validity for academic performance. Although Rules does not aim to measure general intelligence comprehensively, it effectively assesses key aspects of fluid reasoning, thereby offering a practical instrument for educational orientation and research contexts.

The present findings contribute to the literature by providing a psychometrically sound, public-domain alternative to traditional, often costly and time-consuming intelligence assessments. Nonetheless, certain limitations should be acknowledged, including the restricted difficulty range and potential risks associated with item exposure. Future research should investigate the development of adaptive testing formats, the extension to broader ability ranges, and the longitudinal predictive validity of Rules across diverse educational contexts.

Overall, Rules represents a valuable addition to the field of cognitive assessment, promoting more accessible, scalable, and equitable evaluation practices.

References

Abad, F. J., Colom, R., Rebollo, I., & Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: evidence for bias. *Personality And Individual Differences*, *36*(6), 1459–1470. https://doi.org/10.1016/s0191-8869(03)00241-1

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal Of Experimental Psychology: Applied*, *15*(2), 163–181. https://doi.org/10.1037/a0015719

Anum, A. (2022). Does Socio-Economic Status Have Different Impact on Fluid and Crystallized Abilities? Comparing Scores on Raven's Progressive Matrices, Kaufman Assessment Battery for Children II Story Completion and Kilifi Naming Test Among Children in Ghana. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.880005

Avvisati, F. (2020). The measure of socio-economic status in PISA: a review and some suggested improvements. *Large-scale Assessments in Education, 8*(1). https://doi.org/10.1186/s40536-020-00086-x

Barel, E., & Tzischinsky, O. (2018). Age and Sex Differences in Verbal and Visuospatial Abilities. *Advances in Cognitive Psychology, 14*(2), 51–61. https://doi.org/10.5709/acp-0238-x

Bashaw, W. L., & Anderson, H. E. (1967). A correction for replicated error in correlation coefficients. *Psychometrika*, *32*(4), 435–441. https://doi.org/10.1007/bf02289657

Benisz, M., Willis, J. O., & Dumont, R. (2018). Abuses and Misuses of Intelligence Tests: Facts and Misconceptions. In *The MIT Press eBooks*. https://doi.org/10.7551/mitpress/9780262037426.003.0016

Bloemink, S. (2023, 13 July). Botte hakbijl of emancipatiemotor? De geschiedenis van de IQ-test. *De Groene Amsterdammer*. Accessed on 4 September 2024 from https://www.groene.nl/artikel/botte-hakbijl-of-emancipatiemotor

Bouchard, Jr., T. J. (2014). Genes, Evolution and Intelligence. *Behavior Genetics*, *44*(6), 549–577. https://doi.org/10.1007/s10519-014-9646-x

Braaten, E. B., & Norman, D. (2006). Intelligence (IQ) testing. *Pediatrics in Review*, *27*(11), 403–408. https://doi.org/10.1542/pir.27-11-403

Calamia, M., Markon, K., & Tranel, D. (2013). The Robust Reliability of Neuropsychological Measures: Meta-Analyses of Test–Retest Correlations. *The Clinical Neuropsychologist*, *27*(7), 1077–1105. https://doi.org/10.1080/13854046.2013.809795

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-analytic Studies*. New York: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511571312

Carroll, J. B. (1997). The Three-Stratum Theory of Cognitive Abilities. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 122-130). New York: Guilford Press.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for theREnvironment. *Journal Of Statistical Software*, *48*(6). https://doi.org/10.18637/jss.v048.i06

Colom, R., & Flores-Mendoza, C. E. (2007). Intelligence predicts scholastic achievement irrespective of SES factors: Evidence from Brazil. *Intelligence, 35*(3), 243–251. https://doi.org/10.1016/j.intell.2006.07.008

Colom, R., & García-López, O. (2002). Sex differences in fluid intelligence among high school graduates. *Personality And Individual Differences*, *32*(3), 445–451. https://doi.org/10.1016/s0191-8869(01)00040-x

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public domain measure. *Intelligence, 43*(1), 52-64. doi:10.1016/j.intell.2014.01.004

Correlation between the Wechsler Adult Intelligence Scale and Raven's Progressive Matrices. (2018). https://openpsychometrics.org/info/wais-raven-correlation/

Cronbach, L. J. (1990). *Essentials of psychological testing (5th ed.).* New York: HarperCollins.

Cureton, E. E. (1966). Corrected item-test correlations. *Psychometrika*, *31*(1), 93–96. https://doi.org/10.1007/bf02289461

Demulder, L., Lacante, M., & Donche, V. (2020). Large scale measurements to support students in their transition to higher education: The importance of including a non-cognitive perspective. In E. Braun, R. Esterhazy, & R. Kordts-Freudinger (Eds.), *Research on Teaching and Learning in Higher Education* (pp. 11–20). Waxmann Verlag.

DeRue, D. S., Ashford, S. J., & Myers, C. G. (2012). Learning Agility: In Search of Conceptual Clarity and Theoretical Grounding. *Industrial And Organizational Psychology, 5*(3), 258–279. https://doi.org/10.1111/j.1754-9434.2012.01444.x

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement*, 27(6), 440-458.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105,* 399-412. https://doi.org/10.1111/bjop.12046

Duyck, W. (2023). *Mijn kind, slim kind: waarom lezen en tellen de wereld zullen redden*. Pelckmans

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

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. (2010). Het COTAN-beoordelingssysteem voor de kwaliteit van tests herzien. *De Psycholoog, 45*(9), 48–55. http://dare.uva.nl/personal/record/334214

Flanagan, D. P. (2000). Wechsler-based CHC cross-battery assessment and reading achievement: Strengthening the validity of interpretations drawn from Wechsler test scores. *School Psychology Quarterly, 15*(3), 295–329. https://doi.org/10.1037/h0088789

Flemish Government. (2018, November 7). *Leerlingenkenmerken. Naar gemeente school. Schooljaar 2011 – 2012. [Student Characteristics. By Municipality and School. Academic Year 2011 – 2012.]*. https://data-onderwijs.vlaanderen.be/documenten/bestand.ashx?nr=8851

Fonteyne, L. (2017). *Constructing SIMON : a tool for evaluating personal interests and capacities to choose a post-secondary major that maximally suits the potential.* Ghent University. Faculty of Psychology and Educational Sciences.

Fonteyne, L., De Fruyt, F., Dewulf, N., Duyck, W., Erauw, K., Goeminne, K., Lammertyn, J., Marchant, T., Moerkerke, B., Oosterlinck, T., & Rosseel, Y. (2014). Basic mathematics test predicts statistics achievement and overall first year academic success. *European Journal Of Psychology Of Education*, *30*(1), 95–118. https://doi.org/10.1007/s10212-014-0230-9

Fonteyne, L., Duyck, W., & De Fruyt, F. (2017). Program-specific prediction of academic achievement on the basis of cognitive and non-cognitive factors. *Learning And Individual Differences*, *56*, 34–48. https://doi.org/10.1016/j.lindif.2017.05.003

Gignac, G. E. (2018). Conceptualizing and measuring intelligence. In *The SAGE Handbook of Personality and Individual Differences: Volume I: The Science of Personality and Individual Differences* (pp. 439-464). SAGE Publications Ltd.

Gottfredson, L. (1997). Why g matters: The complexity of everyday life. *Intelligence, 24*(1), 79–132. https://doi.org/10.1016/S0160-2896(97)90014-3

Graham, S. (2015). Inaugural Editorial for the Journal of Educational Psychology. *Journal of Educational Psychology, 107*(1), 1-2. https://doi.org/10.1037/edu0000007

Harris, K. R. (2003). Editorial: Is the work as good as it could be? *Journal of Educational Psychology, 95*(3), 451-452. https://doi.org/10.1037/0022-0663.95.3.451

Heeren, J., Speelman, D., & De Wachter, L. (2020). A practical academic reading and vocabulary screening test as a predictor of achievement in first-year university students: implications for test purpose and use. *International Journal Of Bilingual Education And Bilingualism*, *24*(10), 1458–1473. https://doi.org/10.1080/13670050.2019.1709411

Horn, J. L. (1988). Thinking about human abilities. In J. R. Nesselroade & R. B. Cattell (Eds.) *Handbook of Multivariate Experimental Psychology* (pp. 645–685). New York: Academic Press.

Horn, J. L. (1991). Measurement of Intellectual Capabilities: A Review of Theory. In K. S. Mcgrew, J. K. Werder, & R. W. Woodcock (Eds.), *Woodcock-Johnson Technical Manual* (pp. 197-232). Chicago, IL: Riverside.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53–91). New York: Guilford Press.

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. https://doi.org/10.1080/10705519909540118

Hulstijn, J. H. (2015). Language Proficiency in Native and Non-native Speakers. In *Language learning and language teaching*. https://doi.org/10.1075/lllt.41

Immekus, J. C., Snyder, K. E., & Ralston, P. A. S. (2019). Multidimensional Item Response Theory for Factor Structure Assessment in Educational Psychology Research. *Frontiers in Education*, *4*. https://doi.org/10.3389/feduc.2019.00045

Johnson, W., & Bouchard, Jr., T. J. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*(4), 393–416. https://doi.org/10.1016/j.intell.2004.12.002

Kajonius, P. J. (2014). Honesty–Humility in contemporary students: Manipulations of self-image by inflated IQ estimations. *Psychological Reports, 115*(1), 311–325. https://doi.org/10.2466/17.04.pr0.115c13z8

Kaufman, A. S., Raiford, S. E., & Coalson, D. L. (2016). *Intelligent Testing with the WISC-V*. John Wiley & Sons.

Kirkegaard, E. O., & Nordbjerg, O. (2015). Validating a Danish translation of the international cognitive ability resource sample test and cognitive reflection test in a student sample. *Open Differential Psychology*, 1-8.

Kristjánsdóttir & Zaiter (2023). *Public Domain Intelligence Tests: Psychometric properties of the Cog15 and ICAR16 cognitive ability scales* [Master's Thesis, Lund Universiteit]. LUND UNIVERSITY LIBRARIES. https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9125503&fileOId=9170746

Kruglova, N. & Dykhovychnyi, O. (2022). *Choosing MIRT Model for Analysis of Quality of Pedagogical and Psychological Tests.* IEEE 3rd International Conference on System Analysis & Intelligent Computing (SAIC), Kyiv, Ukraine, 2022, 1-4. 10.1109/SAIC57818.2022.9922918.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology, 86*(1), 148–161. https://doi.org/10.1037/0022-3514.86.1.148

Levine, S. Z. (2011). Elaboration on the association between IQ and parental SES with subsequent crime. *Personality And Individual Differences*, *50*(8), 1233–1237. https://doi.org/10.1016/j.paid.2011.02.016

Luo, M., Sun, D., Zhu, L., & Yang, Y. (2021). Evaluating scientific reasoning ability: Student performance and the interaction effects between grade level, gender, and academic achievement level. *Thinking Skills And Creativity*, *41*, 100899. https://doi.org/10.1016/j.tsc.2021.100899

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847–862. https://doi.org/10.3758/brm.42.3.847

Magis, D., Beland, S., & Raiche, G. (2020). Package 'difR'. Collection of Methods to Detect Dichotomous Differential Item Functioning (DIF). Retrieved from https://cran.r-project.org/web/packages/difR/difR.pdf

Marks, G. N., & O'Connell, M. (2021a). No evidence for cumulating socioeconomic advantage. Ability explains increasing SES effects with age on children's domain test scores. *Intelligence, 88*. https://doi.org/10.1016/j.intell.2021.101582

Marks, G. N., & O'Connell, M. (2021b). Inadequacies in the SES–Achievement model: Evidence from PISA and other studies. *Review of Education, 9*(3). https://doi.org/10.1002/rev3.3293

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum.

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

McGrew, K. S., Flanagan, D. P., Keith, T. Z., & Vanderwood, M. (1997). Beyond g: The impact of Gf–Gc specific cognitive abilities research on the future use and interpretation of intelligence tests in the school. *School Psychology Review, 26*(2), 189−210.

McGrew, K. S., & Wendling, B. J. (2010). Cattell–Horn–Carroll cognitive-achievement relations: What we have learned from the past 20 years of research. *Psychology in the Schools, 47*(7), 651–675. https://doi.org/10.1002/pits.20497

McLeod, J. W. H., & McCrimmon, A. W. (2021). Test Review: Raven's 2 Progressive Matrices, Clinical Edition (Raven's 2). *Journal of Psychoeducational Assessment, 39*(3), 388–392. https://doi.org/10.1177/0734282920958220

McLeod, H. N., & Rubin, J. (1962). Correlation between Raven Progressive Matrices and the WAIS. *Journal of Consulting Psychology, 26*(2), 190–191. https://doi.org/10.1037/h0040278

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Pearson (2020). *Raven's 2, Wereldwijd de meest gebruikte non-verbale intelligentietest.* Accessed on 4 September 2024 from https://www.pearsonclinical.nl/pub/media/productfile/b/r/brochure_ravens2_digitaal.pdf

Pearson (2023). *Raven's 2 NL Scores invoeren*. Accessed on 4 September 2024 from https://qglobal.pearsonclinical.com/qg/static/Product/nl/index.htm#Ravens2-NL/Ravens2-NL_Enter_Scores.htm

Peng, P., Wang, T., Wang, C., & Lin, X. (2019). A meta-analysis on the relation between fluid intelligence and reading/mathematics: Effects of tasks, age, and social economics status. *Psychological Bulletin, 145*(2), 189–236. https://doi.org/10.1037/bul0000182

Piraksa, C., Srisawasdi, N., & Koul, R. (2014). Effect of Gender on Student's Scientific Reasoning Ability: A Case Study in Thailand. *Procedia - Social And Behavioral Sciences*, *116*, 486–491. https://doi.org/10.1016/j.sbspro.2014.01.245

Primi, R., Ferrão, M. E., & Almeida, L. S. (2010). Fluid intelligence as a predictor of learning: A longitudinal multilevel approach applied to math. *Learning and Individual Differences, 20*(5), 446–451. https://doi.org/10.1016/j.lindif.2010.05.001

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 4. The Advanced Progressive Matrices.* Harcourt Assessment.

Raven, J.C. & Raven, J. (2018). *Raven's 2 Progressive Matrices Clinical Edition; Nederlandstalige bewerking.* Amsterdam: Pearson Benelux B.V.

Raven, J., Rust, J., Chan, F., & Zhou, X. (2018). *Raven's 2 progressive matrices, clinical edition* (Raven's 2). Pearson.

Reise, S. P., & Haviland, M. G. (2024). Understanding Alpha and Beta and Sources of Common Variance: Theoretical Underpinnings and a Practical Example. *Journal Of Personality Assessment*, 1–16. https://doi.org/10.1080/00223891.2024.2420175

Revelle, W. (2024). psych:Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, https://CRAN.r-project.org/package=psych, 2.4.1 edition. R package version 2.4.1.

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, *138*(2), 353–387. https://doi.org/10.1037/a0026838

Rindermann, H., Becker, D., & Coyle, T. R. (2020). Survey of expert opinion on intelligence: Intelligence research, experts' background, controversial issues, and the media. *Intelligence*, *78*, 101406. https://doi.org/10.1016/j.intell.2019.101406

Rindermann, H., Flores-Mendoza, C., & Mansur-Alves, M. (2010). Reciprocal effects between fluid and crystallized intelligence and their dependence on parents' socioeconomic status and education. *Learning And Individual Differences*, *20*(5), 544–548. https://doi.org/10.1016/j.lindif.2010.07.002

Roth, B., Becker, N., Romeyke, S., Michael, T., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, *53*, 118–137. https://doi.org/10.1016/j.intell.2015.09.002

Author et al. (2022).

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 99–144). The Guilford Press.

Schneider, W. J., & Newman, D. A. (2015). Intelligence is multidimensional: Theoretical review and implications of specific cognitive abilities. *Human Resource Management Review*, *25*(1), 12–27. https://doi.org/10.1016/j.hrmr.2014.09.004

Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. *The American Journal of Psychology, 15*, 201-293. https://doi.org/10.2307/1412107

Sternberg, R. J., & Kaufman, S. B. (Eds.). (2011). *The Cambridge handbook of intelligence.* Cambridge University Press. https://doi.org/10.1017/CBO9780511977244

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch Trees: A New Method for Detecting Differential Item Functioning in the Rasch Model. *Psychometrika, 80*(2), 289–316.

Taber, K. S. (2017). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Trapp, S., & Ziegler, M. (2019). How Openness Enriches the Environment: Read More. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.01123

Vernon, P. E. (1965). Ability factors and environmental influences. *American Psychologist*, *20*(9), 723–733. https://doi.org/10.1037/h0021472

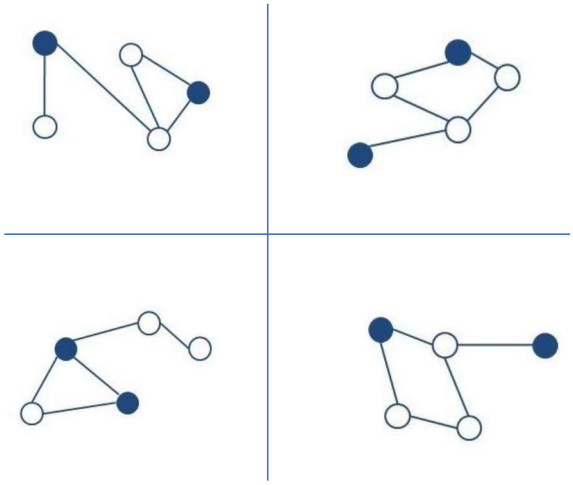
Waschl, N., & Burns, N. R. (2020). Sex differences in inductive reasoning: A research synthesis using meta-analytic techniques. *Personality And Individual Differences*, *164*, 109959. https://doi.org/10.1016/j.paid.2020.109959

Young, S. R., & Keith, T. Z. (2020). An Examination of the Convergent Validity of the ICAR16 and WAIS-IV. *Journal Of Psychoeducational Assessment*, *38*(8), 1052–1059. https://doi.org/10.1177/0734282920943455

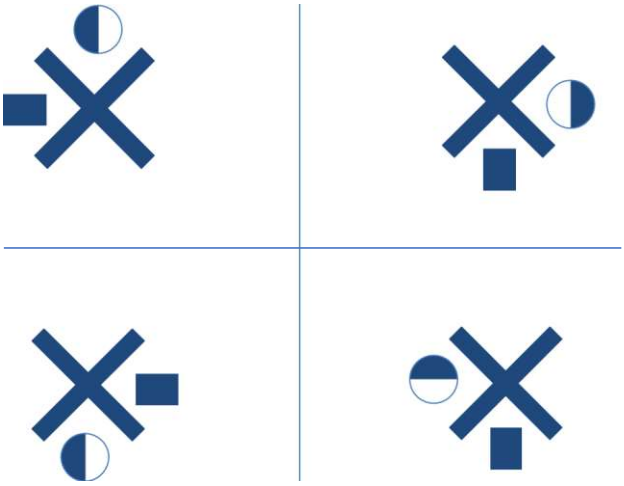
Zinbarg, R., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ωH: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123–133. https://doi.org/10.1007/s11336-003-0974-7

Appendix A

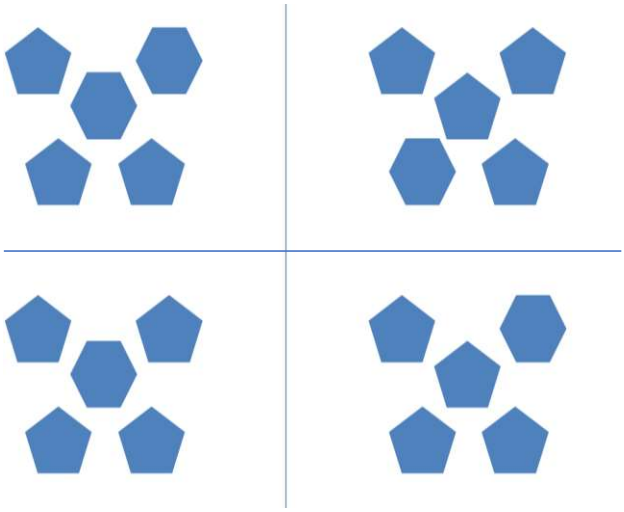
DEDUCTION1 (below left)



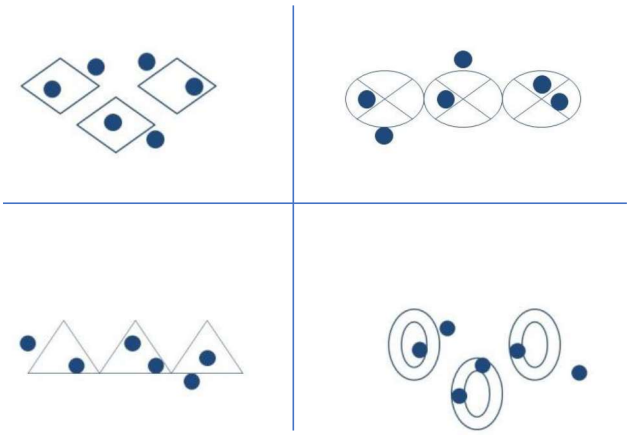
DEDUCTION2 (top right)



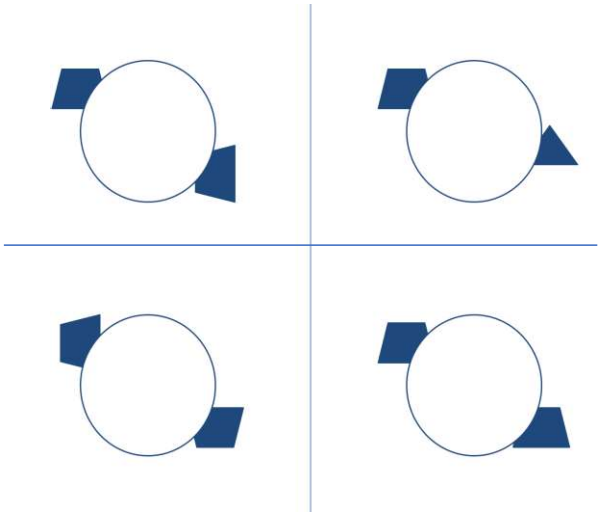
DEDUCTION3 (top left)



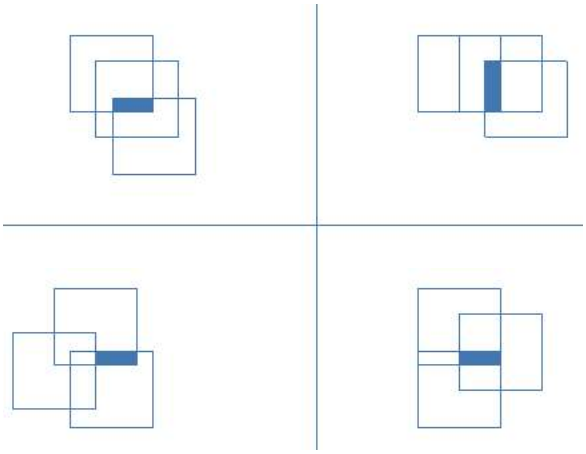
DEDUCTION4 (top left)



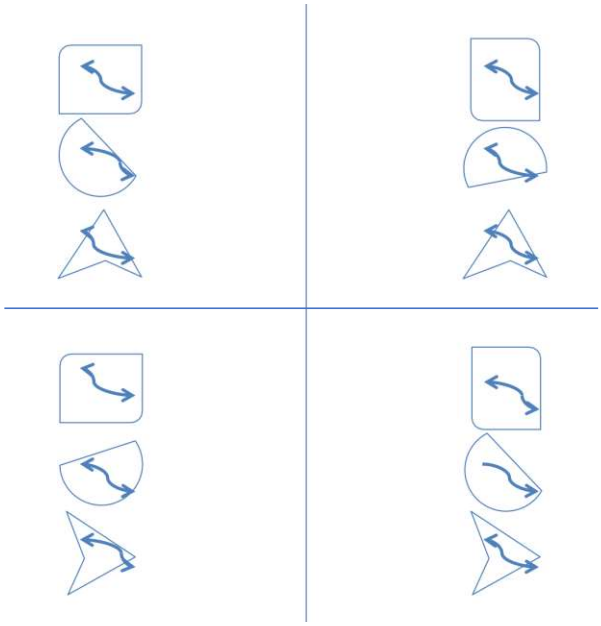
DEDUCTION5 (top right)



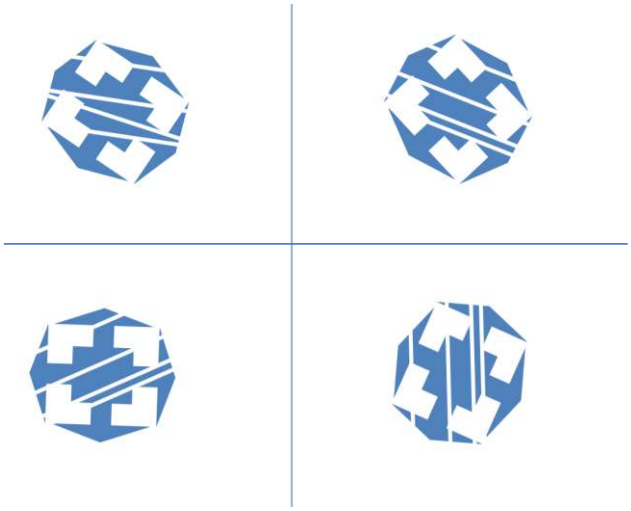
DEDUCTION6 (below left)



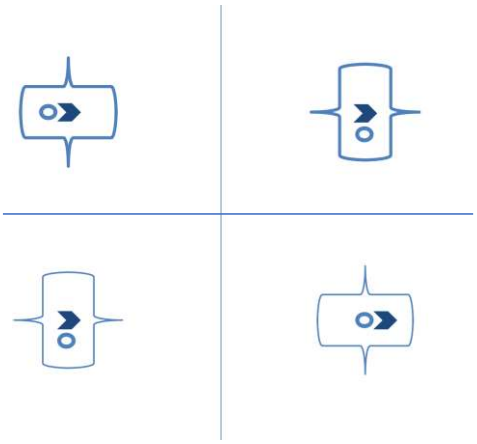
DEDUCTION7 (below right)



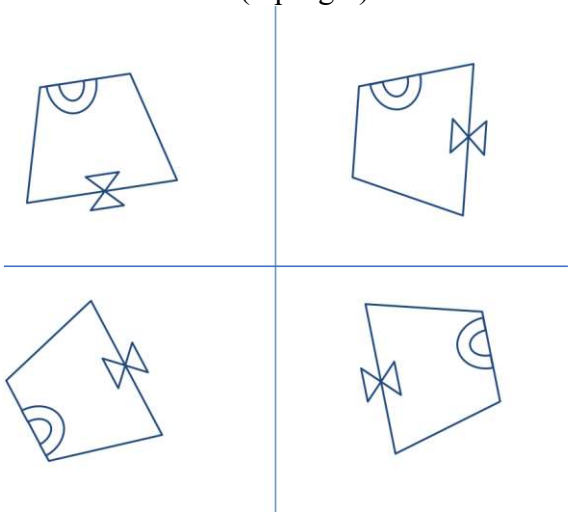
DEDUCTION8 (top left)



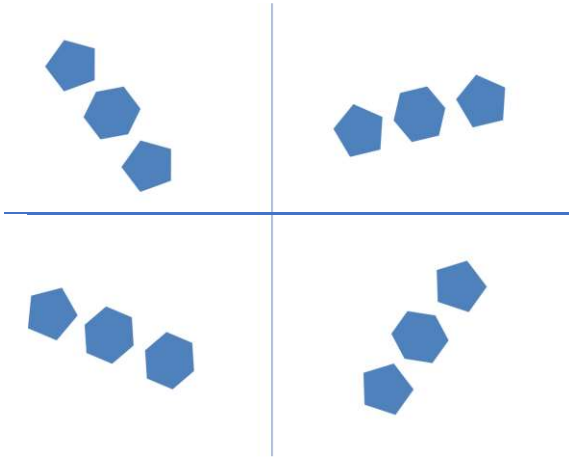
DEDUCTION9 (below right)



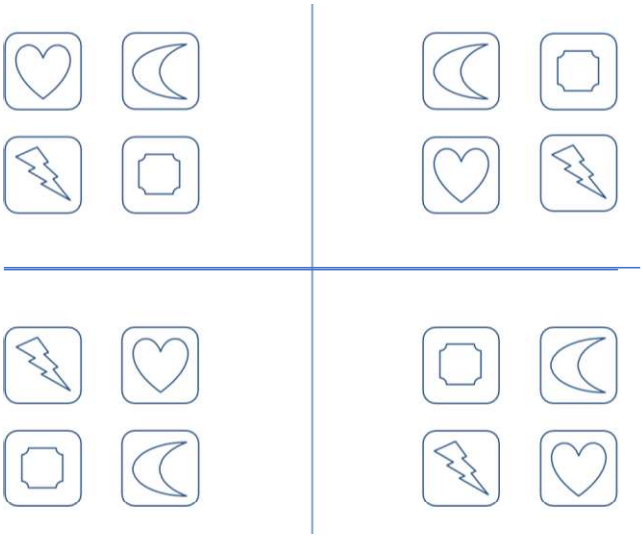
DEDUCTION10 (top right)



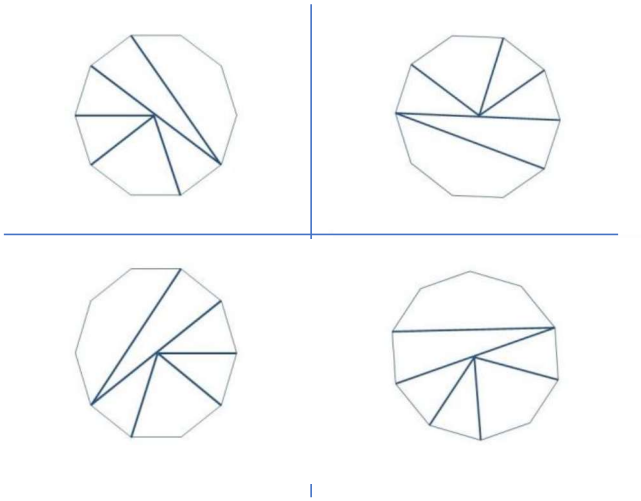
DEDUCTION11 (below left)



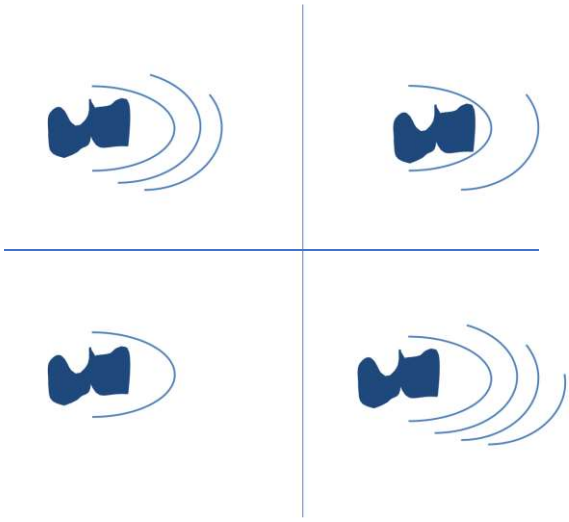
DEDUCTION12 (below right)



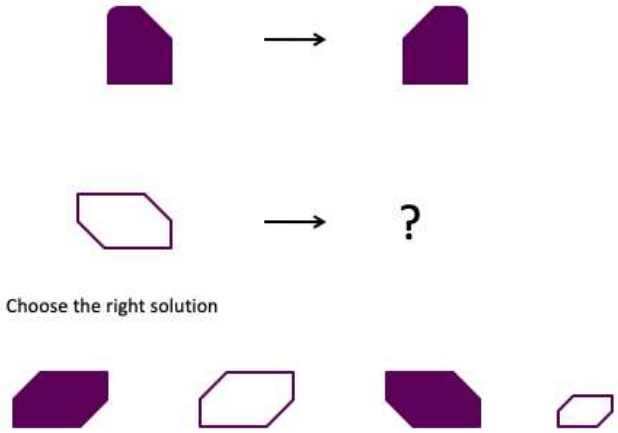
DEDUCTION13 (below left)



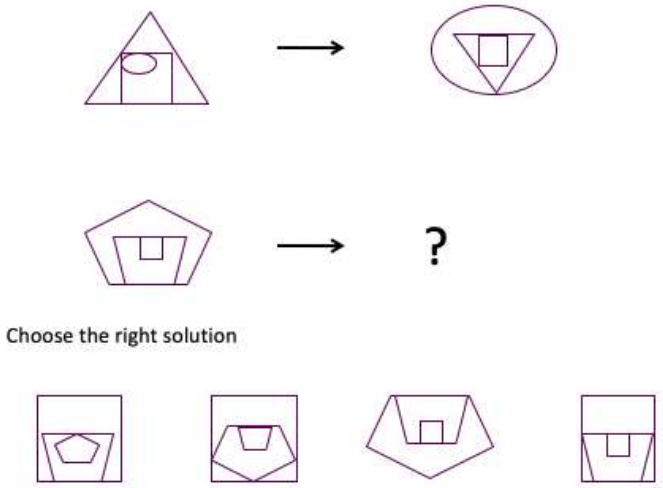
DEDUCTION14 (top right)



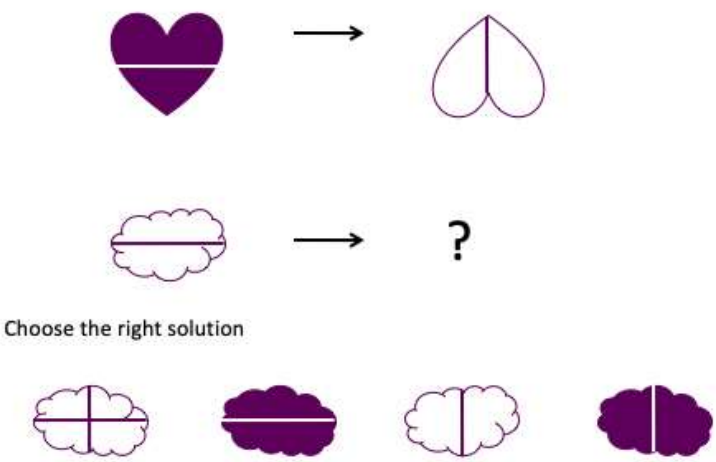
INDUCTION1 (option 2)



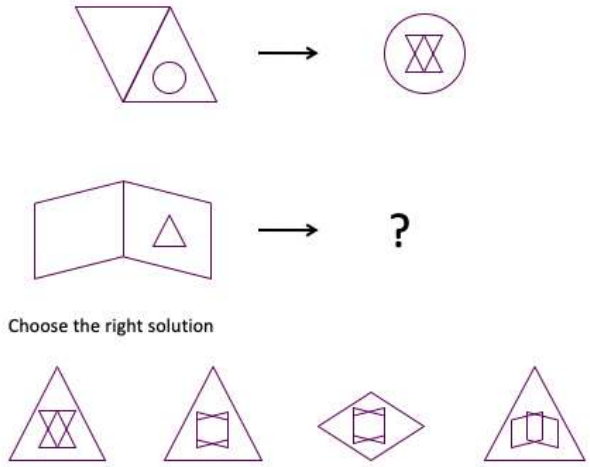
INDUCTION2 (option 2)



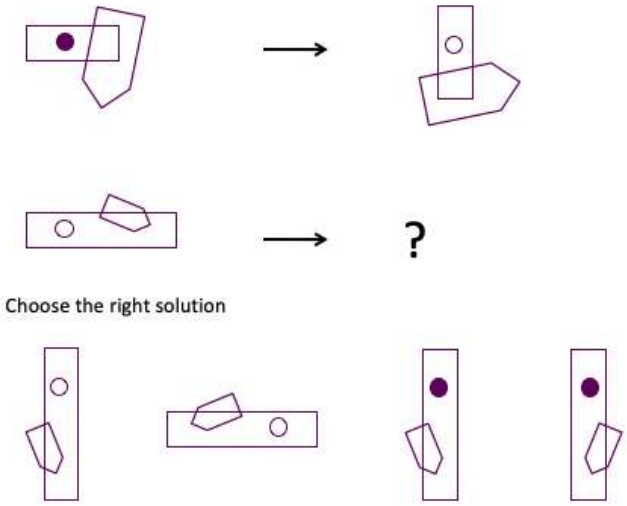
INDUCTION3 (option 4)



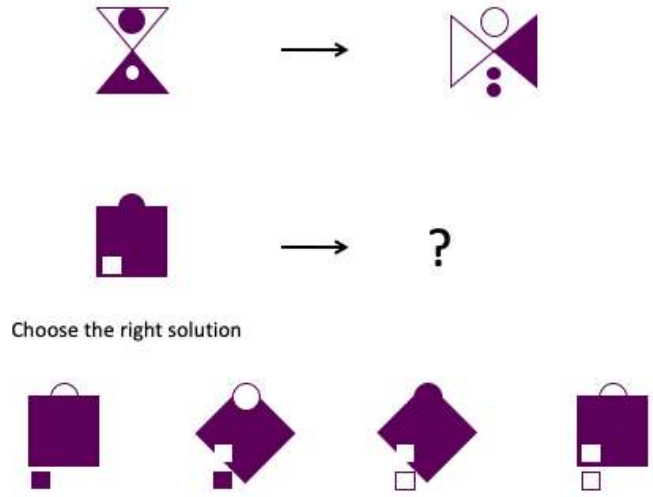
INDUCTION4 (option 2)



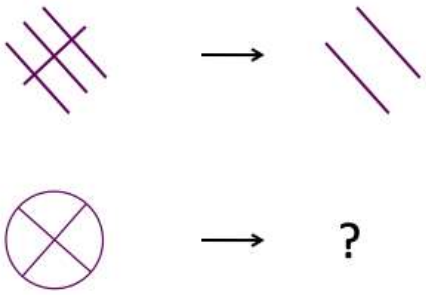
INDUCTION5 (option 3)



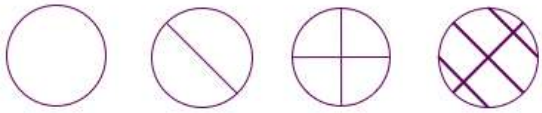
INDUCTION6 (option 1)



INDUCTION7 (option 1)



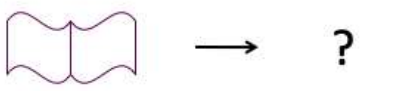
Choose the right solution



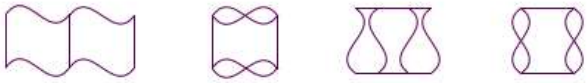
INDUCTION8 (option 2)

Choose the right solution

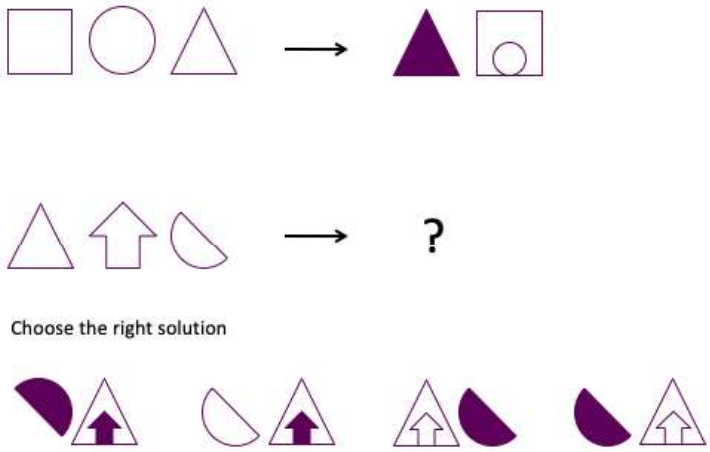
INDUCTION9 (option 2)



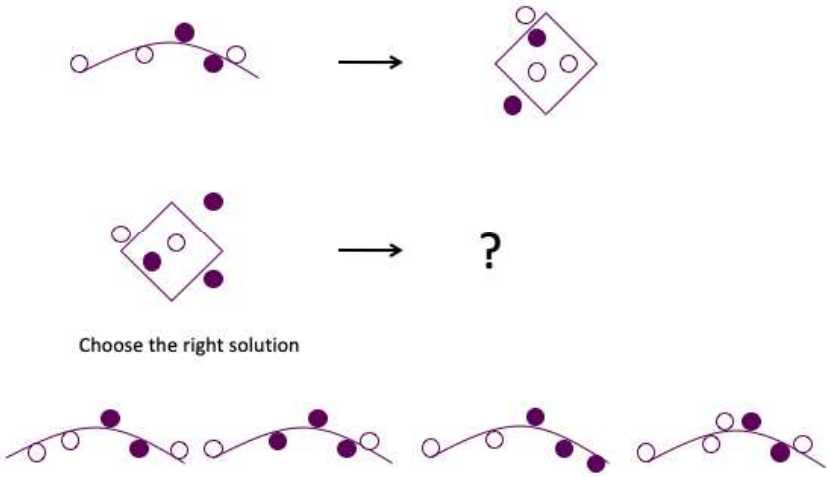
Choose the right solution



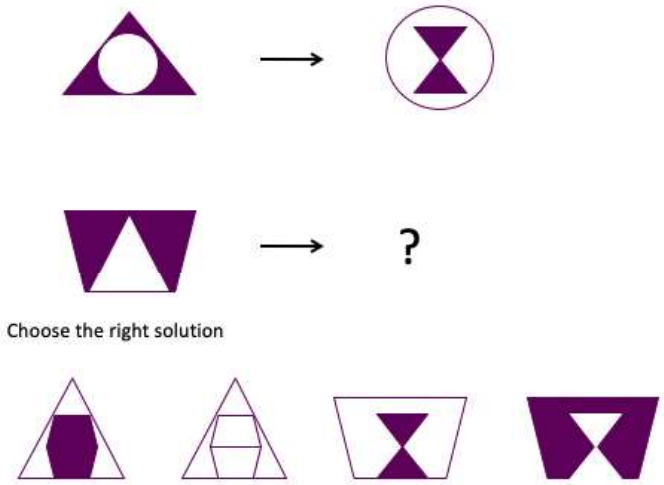
INDUCTION10 (option 4)



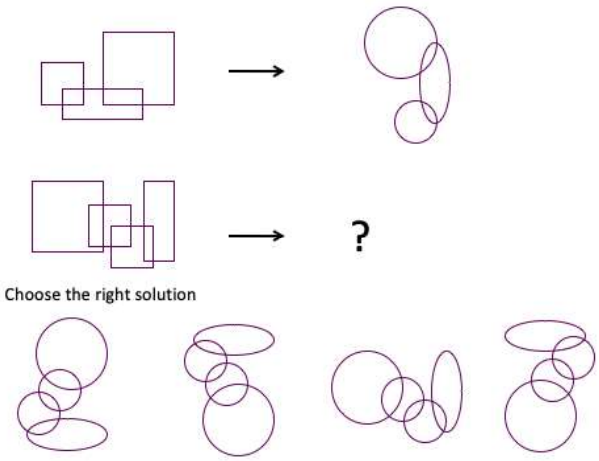
INDUCTION11 (option 3)



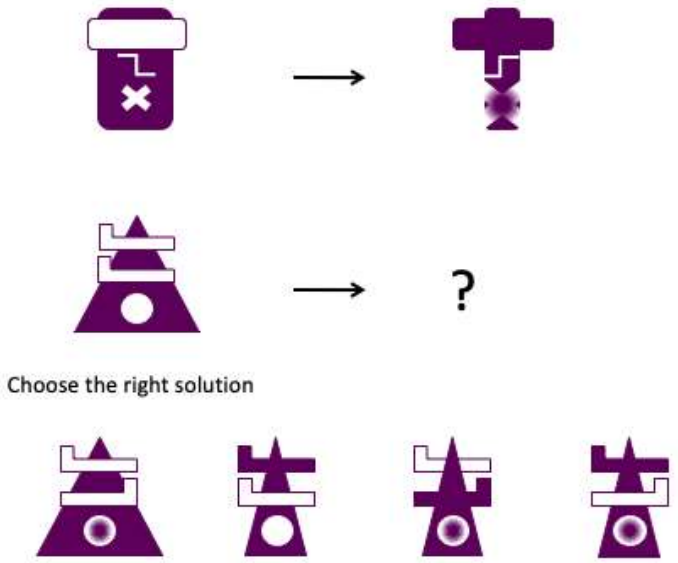
INDUCTION12 (option 1)



INDUCTION13 (option 4)



INDUCTION14 (option 4)



Appendix B

Table B.1

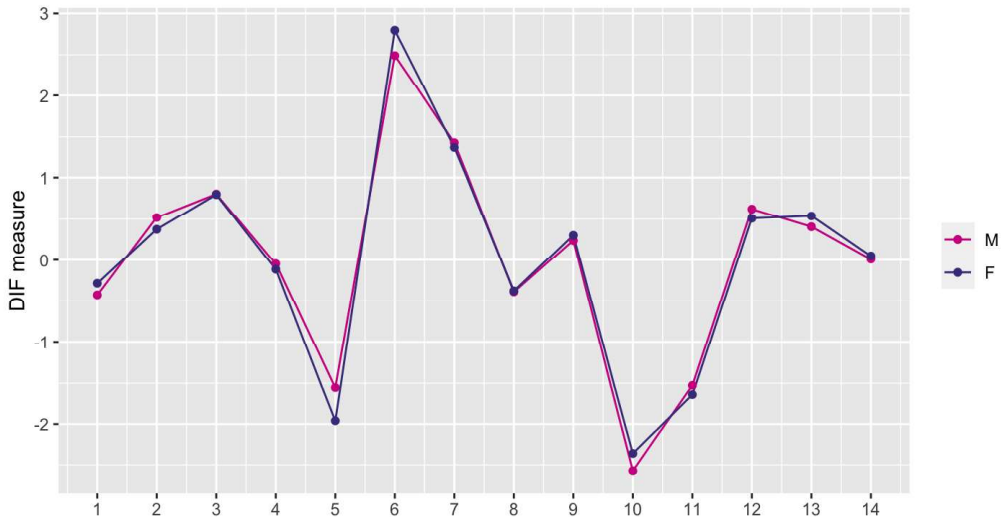
Differential Item Functioning Statistics for Sex

	Mantel-Haenszel Chi-square statistic	P-value	alphaMH	deltaMH	Effect size
Ded1	32.30***	0.00	1.19	-0.41	A
Ded2	30.47***	0.00	0.86	0.35	A
Ded3	0.13	0.71	0.99	0.02	A
Ded4	6.49*	0.01	0.93	0.18	A
Ded5	87.13***	0.00	0.64	1.04	B
Ded6	132.69***	0.00	1.43	-0.84	A
Ded7	0.10	0.76	0.99	0.02	A
Ded8	1.33	0.25	0.96	0.09	A
Ded9	7.09**	0.01	1.08	-0.17	A
Ded10	8.05**	0.00	1.21	-0.46	A
Ded11	14.19***	0.00	0.84	0.41	A
Ded12	19.60***	0.00	0.89	0.28	A
Ded13	9.32**	0.00	1.09	-0.20	A
Ded14	0.16	0.68	0.99	0.03	A
Ind1	74.97***	0.00	0.73	0.74	A
Ind2	74.29***	0.00	0.76	0.64	A
Ind3	9.28**	0.00	0.90	0.25	A
Ind4	37.59***	0.00	1.18	-0.40	A
Ind5	22.47***	0.00	1.15	-0.33	A
Ind6	11.12***	0.00	1.09	-0.21	A
Ind7	0.08	0.78	0.99	0.02	A
Ind8	33.64***	0.00	0.77	0.61	A
Ind9	4.76*	0.03	1.10	-0.22	A
Ind10	0.01	0.93	1.01	-0.01	A
Ind11	10.32**	0.00	1.09	-0.20	A
Ind12	11.99***	0.00	0.89	0.28	A
Ind13	2.53	0.11	1.05	-0.12	A
Ind14	6.53*	0.01	1.07	-0.17	A

Note. The absolute value of deltaMH serves as an effect size indicator for DIF, with classification following the ETS criteria: a negligible effect when $|\Delta MH| \leq 1$ (Class A), a moderate effect when $1 < |\Delta MH| \leq 1.5$ (Class B), and a large effect when $|\Delta MH| > 1.5$ (Class C) (Magis et al., 2020). *** p <.0001, ** $p < 0.001$, * $p < 0.05$, N = 32,585

Figure B.1

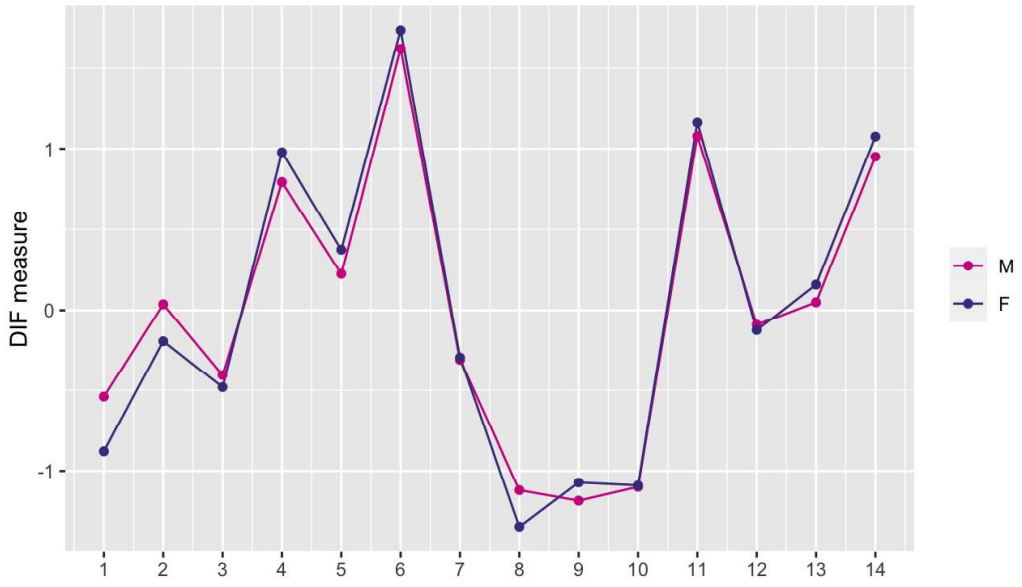
Deduction Differential Item Functioning Statistics for Sex



Note. N = 32,585.

Figure B.2

Induction Differential Item Functioning Statistics for Sex



Note. N = 32,585.

Table B.2

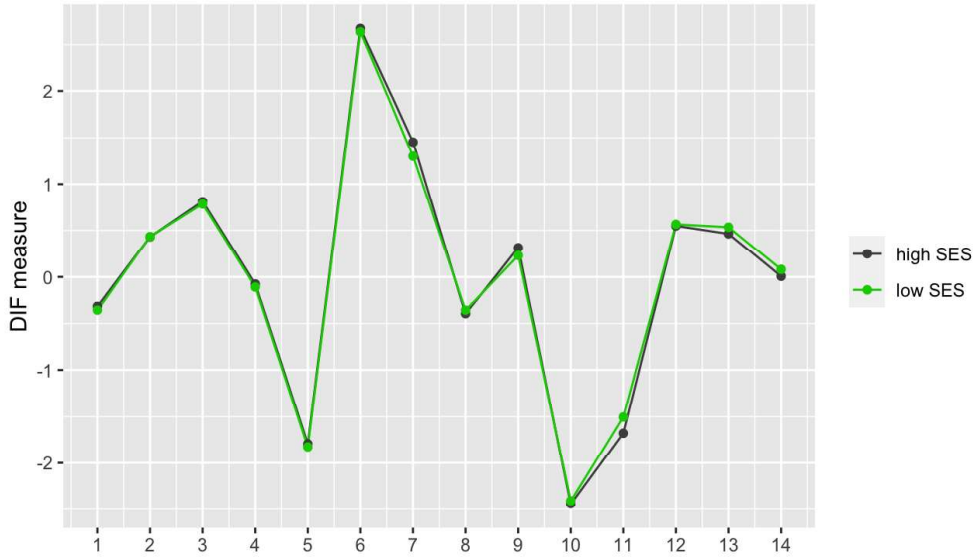
Differential Item Functioning Statistics for SES

	Mantel-Haenszel Chi-square statistic	P-value	alphaMH	deltaMH	Effect size
Ded1	0.21	0.65	0.99	0.03	A
Ded2	1.39	0.24	1.03	-0.07	A
Ded3	0.21	0.65	0.99	0.03	A
Ded4	0.89	0.35	0.97	0.06	A
Ded5	0.91	0.34	0.96	0.11	A
Ded6	0.04	0.84	1.01	-0.02	A
Ded7	11.72***	0.00	0.92	0.20	A
Ded8	0.00	0.95	1.00	-0.01	A
Ded9	5.39*	0.02	0.94	0.15	A
Ded10	0.42	0.52	0.96	0.10	A
Ded11	7.97**	0.00	1.14	-0.30	A
Ded12	4.26*	0.04	1.06	-0.13	A
Ded13	3.25	0.07	1.05	-0.12	A
Ded14	3.00	0.08	1.05	-0.12	A
Ind1	2.84	0.09	0.94	0.14	A
Ind2	2.30	0.13	0.95	0.11	A
Ind3	0.04	0.84	1.01	-0.02	A
Ind4	4.50*	0.03	0.94	0.14	A
Ind5	1.68	0.20	1.04	-0.09	A
Ind6	0.12	0.73	0.99	0.02	A
Ind7	1.05	0.30	1.04	-0.08	A
Ind8	0.03	0.87	0.99	0.02	A
Ind9	0.66	0.42	1.04	-0.08	A
Ind10	13.87***	0.00	1.17	-0.37	A
Ind11	0.00	0.95	1.00	0.00	A
Ind12	0.82	0.37	0.97	0.07	A
Ind13	0.68	0.41	1.03	-0.06	A
Ind14	0.01	0.92	1.00	-0.00	A

Note. The absolute value of deltaMH serves as an effect size indicator for DIF, with classification following the ETS criteria: a negligible effect when $|\Delta MH| \leq 1$ (Class A), a moderate effect when $1 < |\Delta MH| \leq 1.5$ (Class B), and a large effect when $|\Delta MH| > 1.5$ (Class C) (Magis et al., 2020). *** p <.0001, ** $p < 0.001$, * $p < 0.05$, N = 32,585.

Figure B.3

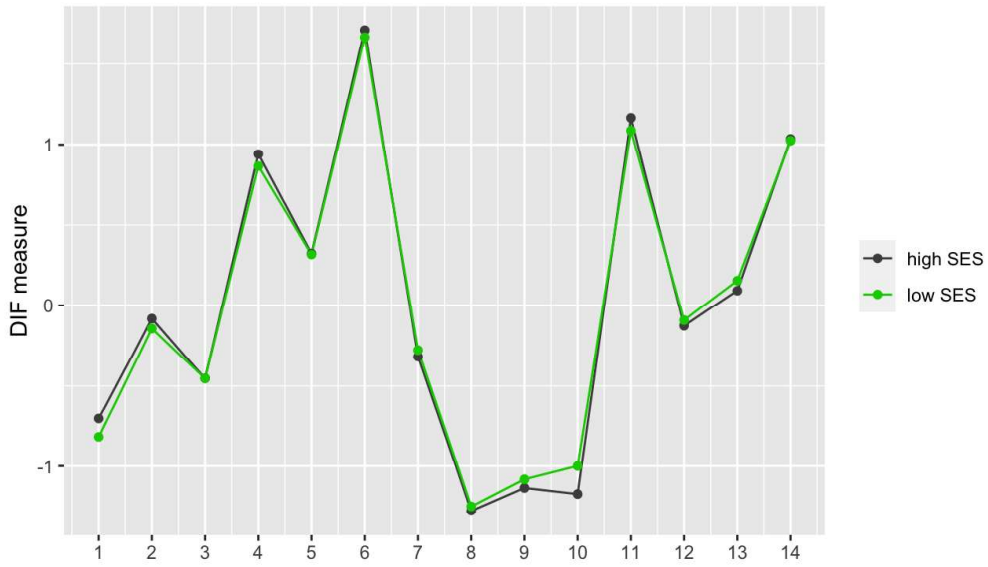
Deduction Differential Item Functioning Statistics for SES



Note. N = 32,585.

Figure B.4

Induction Differential Item Functioning Statistics for SES



Note. N = 32,585.

Table B.3

Factor Loadings MIRT Correlated Variables

	F1	F2	h2
Ded1	0.419	0	0.175
Ded2	0.427	0	0.183
Ded3	0.451	0	0.203
Ded4	0.469	0	0.220
Ded5	0.514	0	0.264
Ded6	0.522	0	0.272
Ded7	0.535	0	0.286
Ded8	0.565	0	0.319
Ded9	0.569	0	0.324
Ded10	0.589	0	0.347
Ded11	0.609	0	0.371
Ded12	0.631	0	0.398
Ded13	0.636	0	0.405
Ded14	0.642	0	0.412
Ind1	0	0.428	0.183
Ind2	0	0.520	0.270
Ind3	0	0.535	0.286
Ind4	0	0.577	0.333
Ind5	0	0.580	0.337
Ind6	0	0.583	0.340
Ind7	0	0.583	0.340
Ind8	0	0.599	0.359
Ind9	0	0.603	0.364
Ind10	0	0.611	0.373

Ind11	0	0.631	0.398
Ind12	0	0.688	0.474
Ind13	0	0.692	0.478
Ind14	0	0.741	0.549

Note. Correlation between F1 and F2 is $r = .885$. $N = 32{,}585$.

Table B.4

Factor Loadings MIRT Second Order Model

	g	S1	S2	h2
Ded1	0.369	0.129	0	0.153
Ded2	0.377	0.132	0	0.160
Ded3	0.391	0.137	0	0.172
Ded4	0.408	0.143	0	0.187
Ded5	0.441	0.154	0	0.218
Ded6	0.447	0.156	0	0.224
Ded7	0.456	0.159	0	0.234
Ded8	0.480	0.168	0	0.259
Ded9	0.480	0.168	0	0.258
Ded10	0.491	0.172	0	0.270
Ded11	0.505	0.177	0	0.286
Ded12	0.520	0.182	0	0.303
Ded13	0.523	0.183	0	0.307
Ded14	0.527	0.184	0	0.312
Ind1	0.376	0	0.136	0.160
Ind2	0.441	0	0.159	0.220
Ind3	0.454	0	0.164	0.233
Ind4	0.484	0	0.174	0.265

Ind5	0.487	0	0.175	0.267
Ind6	0.487	0	0.176	0.268
Ind7	0.488	0	0.176	0.269
Ind8	0.498	0	0.180	0.281
Ind9	0.503	0	0.181	0.286
Ind10	0.506	0	0.182	0.289
Ind11	0.519	0	0.187	0.304
Ind12	0.553	0	0.199	0.345
Ind13	0.556	0	0.200	0.349
Ind14	0.584	0	0.210	0.385

Note. N = 32,585.

Figure 1

Example of Deduction Item

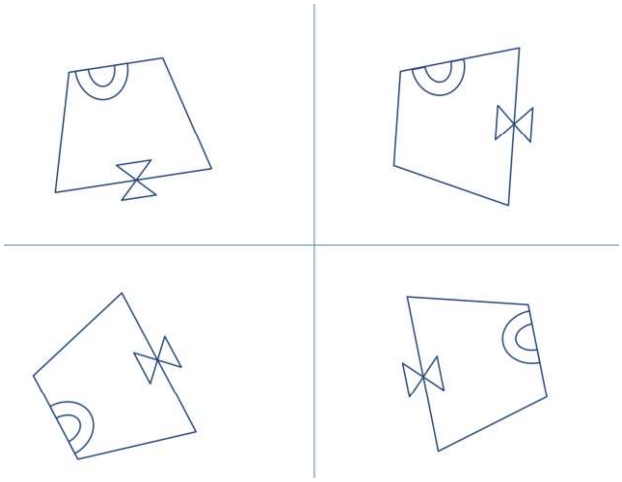


Figure 2

Example of Induction Item

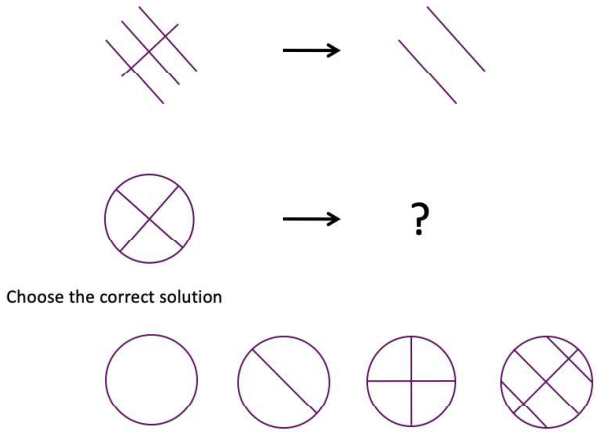


Figure 3

Distribution of Standardized Rules z-scores

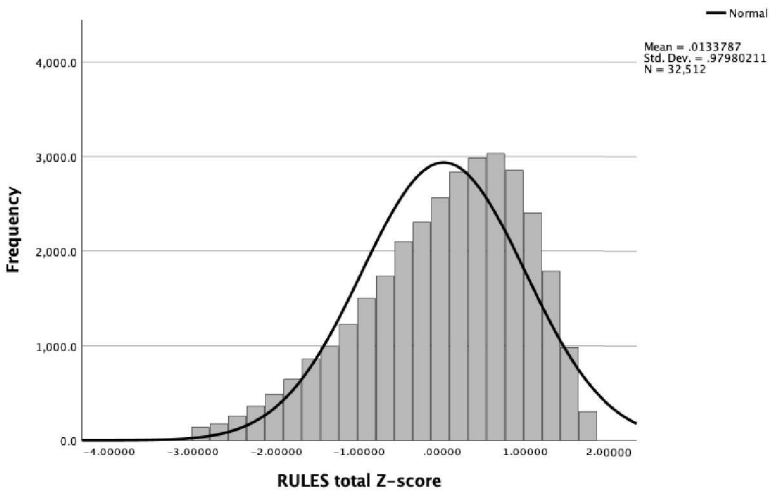
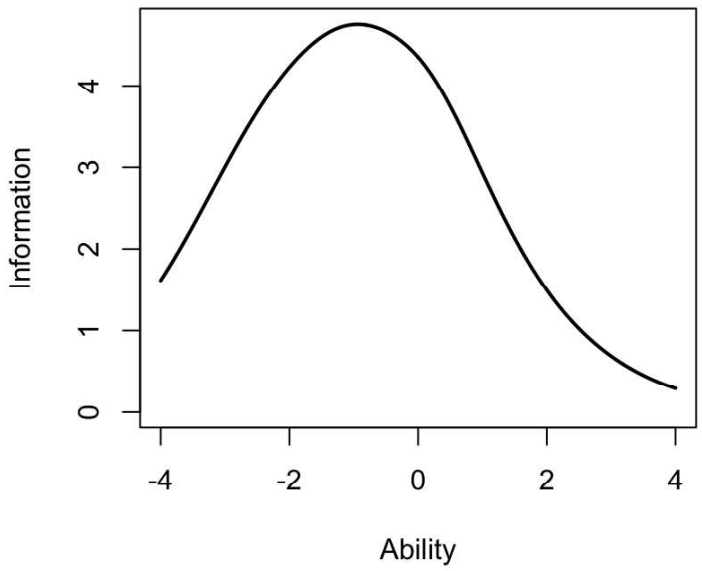


Figure 4

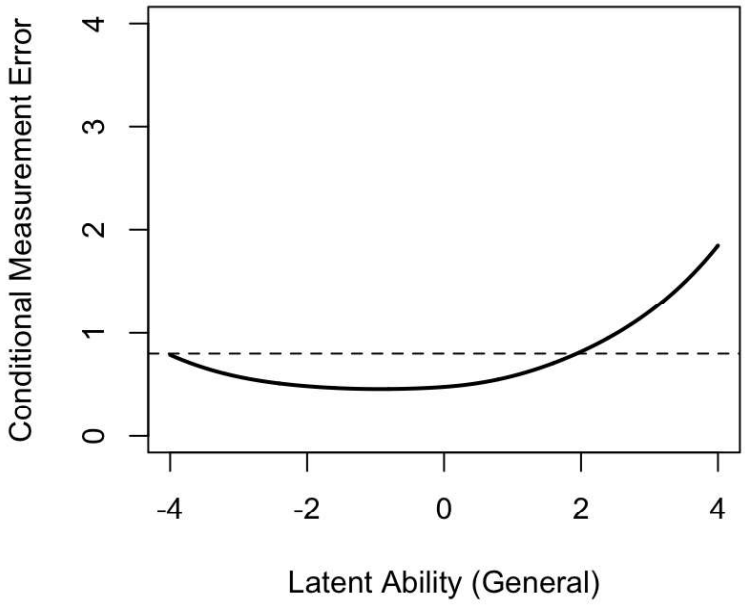
Test Information Function for the 28 Item Rules Non-Verbal Test



Note. 67.3% of the test information lies between -4 and 0. $N = 32{,}585$.

Figure 5

Conditional Reliability for the 28 Item Rules Non-Verbal Test



Note. Conditional reliability coefficients indicate the reliability for subgroups at various cut scores on a test. Rules shows more measurement errors in the high latent ability subgroup. *N* = 32,585.

Declaration of Interest statement

The authors have no relevant financial or non-financial interests to disclose.