Validation of the Children's International Cognitive Ability Resource (Ch-ICAR)

Merel Dutry[a], Alexandra Vereeck[a,b,c], Wouter Duyck[a,d,e], Eva Derous,[e,f] , Stijn Schelfhout[a,e], Arnaud Szmalec[a,g], Evy Woumans[c], Mark Schittekatte[h,], Dries Debeer[i], and Nicolas Dirix[a]

[a]Ghent University, Faculty of Psychology and Educational Sciences, Department of Experimental Psychology, Henri Dunantlaan 2, 9000 Ghent, Belgium.

[b]Ghent University, Faculty of Arts and Philosophy, Department of Linguistics, Blandijnberg 2, 9000 Ghent, Belgium.

[c]Ghent University, Faculty of Arts and Philosophy, Department of Translation, Interpreting and Communication, Groot-Brittanniëlaan 45, 9000 Ghent, Belgium.

[d]The Accreditation Organisation of the Netherlands and Flanders (NVAO), Parkstraat 83, 2514 JG Den Haag, The Netherlands

[e]Ghent University, Faculty of Psychology and Educational Sciences, Department of Work, Organisation, and Society, Vocational and Personnel Psychology Lab (VoPP), Henri Dunantlaan 2, 9000 Ghent, Belgium

[f]Erasmus University Rotterdam, Erasmus School of Social and Behaviourial Sciences, Work and Organizational Psychology, Burgemeester Oudlaan 50, 3062 PA Rotterdam, The Netherlands

[g]Psychological Sciences Research Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium

[h]Ghent University, Assessment Lab of Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

[i]Ghent University, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

Abstract

The International Cognitive Ability Resource, abbreviated ICAR, counters some of the practical problems researchers face when using good, but proprietary licensed intelligence tests like the Wechsler tests, which include unfeasible administration times and financial costs. So far, ICAR has been validated for adolescents and adults in many countries, offering a viable test alternative for these populations. For use among children, however, the appropriateness of this resource was yet unknown. Therefore, we set out to develop a children's ICAR: an instrument composed of ICAR-items, which provides a measure of cognitive ability in children between 11 and 14 years of age. The present article discusses the compilation process of the Ch-ICAR drawing from a pilot study, and evaluates its validity based on two additional studies. The pilot study involved 99 primary school pupils and aimed to select items for the Ch-ICAR instrument. Study 1 investigated the basic psychometric qualities of the Ch-ICAR in a sample of 820 secondary school pupils. Study 2 examined the construct validity by cross-validating the Ch-ICAR with on the one hand Raven's 2 Progressive Matrices, and on the other hand the Flemish CoVaT-CHC Basic Version, relying on samples of 91 secondary and 96 primary school pupils respectively. Results support the utility of the Ch-ICAR as a measure of children's cognitive abilities within a research context.

Keywords: ICAR, cognitive ability, test development, children's cognition

Introduction

"[L]ife is an intelligence test, and the importance of measuring intellectual ability as people respond to life's challenges cannot be overstated." (Dworak et al., 2021)

Cognitive ability refers to the capacity to acquire, process, manipulate, organize and apply knowledge effectively (Gottfredson, 1997). It is widely recognized to reliably predict important outcomes such as academic achievement (Roth et al., 2015; Tikhomirova et al., 2020; Vilia et al., 2017) and work performance (Sackett et al., 2017; Schmidt, 2016). Furthermore, cognitive ability is significantly associated with happiness (Ali et al., 2013; Kanazawa, 2013) and overall health (Wrulich et al., 2014). Unsurprisingly, cognitive ability and its assessment therefore belong to the most discussed and central topics within psychology (Gottfredson & Saklofske, 2009).

For children, the most commonly used test worldwide to measure cognitive ability is the Wechsler Intelligence Scale for Children (WISC; Evers et al., 2012; Flanagan & Alfonso, 2017; Weiss et al., 2019). The WISC is a time- and field-tested instrument that has strong psychometric features (Oakland et al., 2016) and boasts international positive reviews (e.g., Na & Burns, 2016 for the English version; Kwaliteitscentrum voor Diagnostiek, 2023 for the Dutch version). However, the WISC is primarily designed for clinical contexts, and less attuned to the needs of many research settings. Because this test can only be administered individually, the workload soon becomes unmanageable for studies aiming at larger samples, even when using abbreviated scales like the WASI-II (Wechsler, 2011) or other short forms containing a reduced number of subtests (Aubry & Bourdin, 2018). Financially as well, the cost of administering the WISC often surpasses budgets of (junior) researchers.

To alleviate the practical problems that researchers encounter with Wechsler Scales and other proprietary licensed intelligence tests, and "to address the need for psychometrically valid tools that are well-suited for large-scale, remote data collection" (Dworak et al., 2021, p.

3), an interuniversity team of cognitive ability experts developed the International Cognitive Ability Resource (ICAR; Revelle et al., 2020), a public domain measure of cognitive ability designed for online administration.

Public domain measures are valuable tools in a research context. Open access instruments enable more researchers to conduct studies, share findings, and ultimately advance the field (Condon & Revelle, 2014). The public nature of these instruments encourages collaboration within the research community regarding test development, refinement, and validation (Goldberg, 1999). On the downside, publicly available measures of cognitive ability sometimes either lack sufficient psychometric rigor (e.g., Cog15; Kajonius, 2014; Kristjánsdóttir & Zaiter, 2023) or are too narrow in focus, concentrating exclusively on a particular type of item (e.g., only matrix reasoning; Zorowitz et al., 2023).

The ICAR, on the other hand, is psychometrically sound and offers a wide variety of subtests (Revelle et al., 2017) including items of high and low difficulty. The items can be programmed into any software of choice, and the extensive database (The International Cognitive Ability Resource Team, 2014) makes it possible to tailor the instrument to the intended audience or to switch up the items and minimize test-retest effects. In the earliest days of the project, the ICAR consisted of just 60 items, divided over four subtests: Verbal Reasoning, Three-dimensional Rotation, Matrix Reasoning, and Letter-Number Series (Condon & Revelle, 2014). This original ICAR60, as well as a 16-item subset known as the ICAR16, were validated in a sample of 96,958 participants with an age range from 14 to 90 years (Condon & Revelle, 2014). Since its initial validation, the ICAR has established itself as a cognitive ability measure for research on adolescents and adults (used in 79 published studies according to Dworak et al., 2021, and more since), the majority of studies employing the ICAR16.

The appropriateness of an ICAR for children, however, is still unknown. Considering

the dynamic nature of brain development across the lifespan, with rapid changes in early childhood and more gradual changes in adulthood, it is essential to acknowledge that the ICAR items designed for adults may not be adequate for children (Flanagan & McDonough, 2018). For instance, through neurodevelopment, brains become more organized, efficient, and faster at processing information during childhood and adolescence (Lenroot & Giedd, 2006; Low & Cheng, 2006). Moreover, research suggests that age and frontal lobe development play pivotal roles in shaping cognitive ability (Shaw et al., 2006), highlighting the need for assessment tools tailored to the cognitive abilities of children. Therefore, we aimed to compile an instrument based on ICAR items that unites the existing practical advantages of the ICAR with an adequate measurement of the cognitive ability of children between the ages of 11 and 14. As the ICAR project is not based a priori on a particular theoretical framework such as the Cattell-Horn-Carroll (or CHC-) model of intelligence (Schneider & McGrew, 2012), we also looked into the question which broad cognitive abilities are measured by which ICAR subtest. The compilation of a children's ICAR, or Ch-ICAR in brief, not only broadens the scope of the ICAR project: It also addresses a notable gap in the literature, for at the moment there is a lack of psychometrically sound measures of cognitive ability that are freely available, limited in administration time and suited for children.

In this article, we first discuss the pilot study we carried out to compile the Ch-ICAR. Next, we discuss two studies conducted to assess the psychometrics of this newly compiled instrument. Study 1 investigated the basic psychometric qualities of the Ch-ICAR. Study 2 cross-validated the Ch-ICAR with two commercial tests, on the one hand Raven's 2 Progressive Matrices, and on the other hand the Flemish CoVaT-CHC Basic Version. The former is a measure of nonverbal intelligence, which provides an effective estimate of cognitive ability (McLeod & McCrimmon, 2021) and is commonly used and suitable for group administration (Oakland et al., 2016). The latter is a comprehensive measure of

intelligence, whose composition reflects the CHC-model of intelligence (Magez et al., 2015). In Flanders, the CoVaT offers a valid alternative to other intelligence tests, since − unlike Raven's Matrices − it encompasses multiple broad cognitive abilities (Flanagan & McDonough, 2018). Cross-validation of the Ch-ICAR using both Raven and CoVaT instruments provides us with empirical verification about whether the Ch-ICAR is indeed a valid measure for children's cognitive ability.

Pilot study

In an initial phase, we solicited feedback from experts in the field and from children with the target age to assess the suitability of various ICAR item types[1] for constructing a preliminary instrument. This preliminary instrument consisted of four subtests: Verbal Reasoning, Matrix Reasoning, Number Series, and Figural Analogies, with 16, 11, 23 and 20 items respectively (we adopted all items available on the ICAR-website, except for the subtest Number Series, where we selected a mix of easy and more challenging items). Verbal Reasoning contains questions on vocabulary and general knowledge as well as mathematical problems[2]. Matrix Reasoning items consist of eight geometric shapes, placed in a 3-by-3 grid, with one shape missing; participants have to choose the response option that represents the missing ninth shape. Number Series items show a short sequence of numbers; participants have to complete the sequence by entering the next number of the sequence. Figural Analogies items consist of three colored geometric shapes, with the first two shapes standing

[1] Including Verbal Reasoning, Three-dimensional Rotation, Matrix Reasoning, Letter-Number Series, Number Series and Figural Analogies. We decided not to include Three-dimensional Rotation and Letter-Number Series in the preliminary instrument because the former proved unsuitable for the target audience, with half of the children unable to solve any items, and the latter raised concerns about test conditions—specifically, if participants are allowed (or not prevented) from writing down the alphabet, the items become significantly easier.

[2] As such, in our opinion the name of the subtest does not quite cover the content, but for matters of consistency and comparability, we chose to retain the original subtest name.

in a certain relation to each other; participants have to choose the response option that represents the missing fourth shape, which stands in the same relation to the third as the second does to the first. A specimen of each item type can be found in Table A1 (the final compilation of the Ch-ICAR is discussed further and is available on the ICAR website[3]).

Subsequently, we carried out a pilot study with the preliminary instrument. The main objective of this pilot study was to select items for the Ch-ICAR instrument that provide an accurate estimate of cognitive ability, while putting minimal strain on researchers' time and participants' cognitive load. With that goal in mind, we aimed to reduce the number of items to ten at most per subtest.

Five class groups of three different primary schools in Flanders[4] took part in the pilot study (N_{Pilot} = 99, male 41%). All participants were in their last year of primary education (age 11-12). In Flanders, primary education is marked by its comprehensive nature, the absence of ability groups, and open and free accessibility. As a result, Flemish primary education classrooms exhibit a diverse and inclusive pupil population. For practical reasons, each class only took two or three of the four subtests we had previously selected. The combination and order of subtests were assigned at random, the items were administered digitally, and testing took place in the classroom. In line with the recommendation of the ICAR team to take into account possible interruptions due to device malfunctions (Condon & Revelle, 2014), no time limit was imposed.

We noticed that many participants asked for additional explanation about the subtests Matrix Reasoning and Figural Analogies during the test-taking process. To improve the instructions in the final instrument, we added an example item to each of these two subtests.

[3] For now (pre-peer review), the Ch-ICAR can be accessed via this link: https://ap.lc/UXMoM, with this access data: Username: PeerReviewChICAR; Password: 2HE2wiA1. Upon acceptance, the Ch-ICAR will be available via the homepage, users will only need to create a user account.

[4] For the Verbal Reasoning items, we procured the Dutch translation by Fontaine (Nelissen et al., In preparation). With regard to the instructions, we provided each subtest with a written Dutch explanation ourselves.

Seeing that items of limited complexity were scarce in the ICAR database, we designed the example items ourselves (Table A2). Upon completing an example item, participants receive immediate feedback regarding the correctness of the chosen answer and an explanation of why the chosen answer is either correct or incorrect.

For the final item selection, we prioritized items with stronger item-subscale correlations while maintaining a balanced range of difficulty levels. Additionally, we ensured that selected items showed minimal sex differences and avoided disproportionate response times. This approach allowed us to build a diverse and robust final set of items for the Ch-ICAR. Detailed statistics for each subtest can be found in Appendix B. Ultimately, we selected a total of 31 items for the final Ch-ICAR: 8 items for Verbal Reasoning, Number Series and Matrix Reasoning each, and 7 for Figural Analogies (Appendix C contains the final item set). Table 1 provides a concise overview of the primary findings of the pilot study. This table serves as a brief summary of key insights and observations from the preliminary study. For a more comprehensive understanding and detailed item-level information, readers are invited to consult Appendix B.

Table 1

Primary Findings of the Pilot Study

Subtest	*N*	Number of items in preliminary instrument	Cronbach's alpha in preliminary instrument	Average PC (*SD*) in preliminary instrument	Cronbach's alpha in final Ch-ICAR	Average PC (*SD*) in final Ch-ICAR
Verbal Reasoning	46	16	.50	.34 (.14)	.65	.40 (.24)
Matrix Reasoning	66	11	.63	.29 (.20)	.69	.31 (.25)
Number Series	71	23	.88	.35 (.20)	.85	.43 (.31)
Figural Analogies	56	20	.77	.23 (.18)	.83	.33 (.31)

Note. Average PC is the average proportion of correct responses based on the total score of the participants that completed the full subtest.

Study 1: Basic psychometric qualities of the Children's ICAR

Study 1 investigated the basic psychometric qualities of the Ch-ICAR in a large sample of pupils in Flanders. We assessed (1) the internal structure of the test and the psychometric properties of the individual items, (2) the distribution of the Ch-ICAR scores and the distribution of the Ch-ICAR scores over sex (although small sex differences in specific cognitive abilities may exist, the equality in cognitive ability is well established in literature **(**e.g., Deary et al., 2007; Fergusson & Horwood, 1997; Nisbett et al., 2012)), (3) the internal consistency of the different subtests and the Ch-ICAR as a whole, and (4) the construct validity by examining the relation with school achievement and a basic mathematical skill test.

Method

Participants

A sample of Flemish pupils (N_1 = 820, male 53.5%, female 45.7%, .8% did not specify their sex) was recruited in 23 different schools all across Flanders. All participants were in their first year of regular secondary education (M_{age} = 12.02, SD_{age} = .56) and other than that there were no exclusion criteria for participation. Table 2 offers a comprehensive overview of the demographics and associated statistics.

The research project[5] encompassing the ICAR studies was approved by the Ethical Commission of the Faculty of Psychology and Educational Sciences of Ghent University (reference number 2021/59). Both the parents of the participants and the participants themselves gave their informed consent for participation, as well as for storage and use of the data by the researchers.

[5] Entitled: "Study orientation in secondary education".

Table 2

Demographic Information Study 1 and Study 2

	Study 1	Study 2 (RPM sample)	Study 2 (CoVaT sample)
N	820	91	96
N female (%)	375 (46%)	55 (60%)	48 (50%)
Mean age (*SD*) (years)	12.02 (.56)	12.99 (.57)	11.05 (.27)
Age range (years)	10 - 14	12 - 15	10 - 12
Socio-economic status			
N home language is not Dutch (%)	83 (10%)	3 (3%)	3 (3%)
*N r*eceiving a bursary (%)	231 (28%)		
N highest attained degree is primary education or lower Parent 1 (%)	83 (10%)		
N highest attained degree is primary education or lower Parent 2 (%)	81 (10%)		
N schools	23	2	3
M number of pupils per school *(SD)*	35.65 (22.95)	45.50 (38.89)	32.00 (23.64)
Number of pupils per school: range	8- 83	18 - 73	15 - 59

Materials

Socio-economic status. To define the socio-economic status (SES) of the participants, we used a questionnaire for parents with three indicators that are official markers for school funding in Flanders: financial capacity of the parents, educational level of both parents, and

home language (Ministry of Education and Training, 2012). Following official criteria for these indicators, the following participant characteristics were seen as indicative for low SES: receiving a bursary, having one or two parents whose highest attained degree is primary education or lower, and having another home language than Dutch. In our sample, 10% of the participants had another home language than Dutch (for 8% home language was unspecified), 28% of the parents received a bursary (for 17% financial capacity was unspecified), 10% had one parent with primary education or lower as highest degree and 5% had both parents with primary education or lower as highest degree (for 9% educational level of the parents was unspecified).

Children's ICAR. To recapitulate, the Ch-ICAR as compiled after our pilot study comprises eight Verbal Reasoning items, eight Matrix Reasoning items, eight Number Series items, and seven Figural Analogies items (see Pilot study for a description of the subtests). Except for Number Series items (which is open-ended), each item consists of eight response options, including *'none of these'* and *'I don't know'*. Responses on the items are converted to binary values: Right answers are scored as one, wrong answers as zero. The subtest scale scores are calculated as the sum of the binary values of each item of the subtest. The Ch-ICAR total score is calculated as the sum of the subtest scale scores.

Academic performance. ***School achievement.*** Data on school achievement include the GPA (weighted average of assignments and tests of all subjects on a scale from 0 to 100), as well as four separate subjects scores: Mathematics, Dutch, Natural Sciences, and Technics (also on a scale from 0 to 100). As secondary education in Flanders does not feature standardized tests at the time of writing, the GPA and subject scores are based on the assessment of the individual teachers. To avoid possible bias, we standardized the achievement scores within schools[6]. Schools with less than ten participants were excluded for

[6]The first year of secondary education in Flanders consists of two primary streams: 1A and 1B (Ministry of Education and Training, s.d.). Admission to 1A is automatic for those who possess a certificate of primary

the corresponding analyses as well as pupils that changed schools within the year. This resulted in a sample of 722 pupils, drawn from a total of 19 distinct schools.

Basic mathematical skill test: CDR. We included an additional, standardized external criterion, related to cognitive ability, to detect and control for school specific effects. We administered a basic mathematical skill test to all participants, namely the shortened version (18 items, McDonald's omega (ω) = .72) of the *Cognitive Developmental Skills in Arithmetics* (*Cognitieve Deelhandelingen van het Rekenen* [CDR]; Desoete & Roeyers, 2006).

Procedure

Participants completed the Ch-ICAR and the CDR at the beginning of the school year (September-November 2021). The instruments were administered digitally, in the classroom of the participants, during school hours. The pupils' teacher as well as a university supervisor were present. The tests were programmed in Qualtrics (https://www.qualtrics.com), a commercial survey software and research platform.

After a general introduction, participants could work in silence at their own pace on the different tests. The subtests of the Ch-ICAR appeared in a fixed sequence: Verbal Reasoning, Matrix Reasoning, Number Series, and Figural Analogies. The items appeared one by one, in ascending difficulty (as defined in the pilot study). Once participants decided to move on to the next item, they could not go back to previous questions, both for the CDR and the Ch-ICAR. All items required a response. Participants were explicitly told not to ask content-related questions, nor could they use supplementary materials (e.g., calculator, pen

education, while entry into 1B is automatic for those without the certificate. However, pupils holding a certificate of primary education may gain admission to 1B through a favorable decision of the admissions class council and the student guidance center, and with parental agreement. Consequently, the predominant majority of pupils embark on the 1A track (87% in the academic year 2021-2022; Ministry of Education and Training, 2022). As pupils in 1B follow a distinct curriculum, we opted for the separate calculation of standardized scores for this subgroup within the school. B-groups with fewer than 10 pupils were also excluded from the corresponding analyses, resulting in the exclusion of 6 B-groups.

and paper). Most participants finished the Ch-ICAR within 25 minutes[7].

Data on school achievement were communicated to us at the end of the school year (June 2022) by the school principal.

Analyses

Analyses were conducted in SPSS (version 29.0; IMB Corp., 2022) and R (version 4.3.3; R Core Team, 2024). All R analysis codes as well as the variance-covariance matrix are available at: https://osf.io/6cwfs/?view_only=4a80feb83bd94c378d3bd06853820563.

To examine the internal structure of the test and the psychometric properties of individual items, we conducted a hierarchical (aligned with CHC-theory) Item Response Theory analysis (IRT; Chalmers, 2012), using the mirt package in R (version 1.42; Chalmers, 2023). The expected structure based on CHC-theory consists of two levels: several broad cognitive abilities at the first, lowest level (represented by the four Ch-ICAR subtests) and one general[8] cognitive ability at the second, highest level (Flanagan & McDonough, 2018). Items with standardized slopes below 0.20 were removed one at a time, starting with the item with the lowest loading. After each removal, the model was re-estimated. Subsequently, we conducted a Differential Item Functioning (DIF) analysis on the fitted hierarchical IRT model, using score-based structural change tests (Merkle et al., 2014), as implemented in the scDIFtest-package in R (version 0.1.1; Debeer, 2020). Items showing significant DIF, with a p-value below .05 after False Discovery Rate (FDR; Benjamini & Hochberg, 1995) correction, were removed. The final model was re-estimated, and the model fit was evaluated.

To examine the distribution of scores, mean scores and standard deviations per subtest were calculated and a histogram of the Ch-ICAR total score was plotted. For this plot, the raw

[7] Our time registration records encompassed the entire test administration, including the CDR-test and inter-test breaks, rather than specific segments. As a result, we can only provide an approximation of the time needed for completing the Ch-ICAR.

[8] When cognitive ability is explicitly linked to the theoretical framework of the CHC-model, we use the term "general cognitive ability" instead of “cognitive ability” (Flanagan & McDonough, 2018).

total scores were transformed into z-scores ($M = 0$, $SD = 1$). To examine potential sex differences in the total score, we conducted an independent samples t-test with sex as grouping variable and the Ch-ICAR total score as dependent variable.

The internal consistency of the different subtests and the Ch-ICAR as a whole is determined by calculating Cronbach's alpha and McDonald's omega[9] (Dunn et al., 2014; McDonald, 1999; Raykov, 1998) in R using the MBESS package (version 4.9.3; Kelley & Lai, 2012). To assess the reliability coefficients, we adhere to the COTAN guidelines for test quality evaluation (Evers et al., 2010). The COTAN rating system has established cut-off criteria for assessing reliability coefficients, which vary depending on the intended purpose of the test. COTAN distinguishes between three main purposes: tests for important decisions at individual level, tests for relatively less important decisions at individual level and tests for research at group level (Evers et al., 2010, p.34). Since our primary goal is to develop an instrument suitable for research purposes, we apply the cut-off criteria designated for research at group level. According to these criteria, reliability coefficients exceeding .70 are considered good, those ranging between .60 and .70 are deemed sufficient, while coefficients below .60 are deemed insufficient.

To examine the relationship with academic achievement, we calculated Pearson correlations and conducted mixed effects models for the different measures of academic achievement with the lmerTest package (version 3.1.3; Kuznetsova et al., 2017) in R. To assess the unique contribution of the Ch-ICAR total score to the marginal explained variance in academic achievement, we contrasted the marginal explained variance between the full model and a model excluding the Ch-ICAR score (Nakagawa & Schielzeth, 2013; Shaw et al.,

[9] When the assumption of tau-equivalence is violated (which is frequently the case in psychology), omega outperforms alpha (Dunn et al., 2014; Zinbarg et al., 2005). Omega has less risk of overestimation or underestimation of reliability, which makes it the preferred choice.

2023). The marginal explained variance was calculated via the MuMIn package (version 1.47.5; Bartoń, 2022).

Results

Internal structure and psychometric properties

In the initial hierarchical IRT model, with all items included, three items (MR5, MR8, and FA7) demonstrated low standardized slopes (i.e. <0.200), with values of 0.191, 0.199, and 0.110, respectively. We removed FA7, the item with the lowest slope, and re-estimated the model. In the second iteration, two items (MR5 and MR8) still had slopes below 0.200 (0.191 and 0.199, respectively), so MR5 was removed, and the model was estimated again. In the third iteration, MR8 alone exhibited a standardized slope below 0.200 (0.192) and was removed. By the fourth iteration, all remaining items had standardized slopes above the 0.200 threshold.

We conducted a DIF analysis on the fitted hierarchical IRT model (excluding items MR5, MR8 and FA7), which revealed significant DIF only in item NS6, with a Lagrange Multiplier (LM) statistic of 16.510 (df = 2, FDR-corrected p = .007). After removing item NS6, we re-estimated the hierarchical IRT model and evaluated the final model. Although the S-χ^2 statistic is not specifically designed for hierarchical IRT models, we used this statistic to assess item misfit, as the hierarchical structure of the model supports the use of summed item scores. The S-χ^2 test for three items (VR4, VR5, and NS4) yielded FDR-adjusted p-values below .05 (Table 3). However, after visually inspecting the observed response curves, these items were retained (Figure 1, 2, and 3). The S-χ^2 test results for all items are provided in Appendix D.

The standardized and unstandardized slopes for the final model are presented in Table 4, and the DIF statistics are shown in Table 5. The final model exhibited a good overall fit to

the data: RMSEA: .027, 95% CI [.022, .031]; SRMR: .040; CFI: .976; TLI: .973[10]. The estimated regression weights for general cognitive ability on the four dimensions are as follows: Matrix Reasoning = 1.111, Verbal Reasoning = 1.390, Number Series = 1.845, and Figural Analogies = 0.654, indicating that general cognitive ability has a smaller load on the Figural Analogies dimension. Given the improved psychometric properties of this 27-item subset compared to the 31-item pilot version, we used this refined 27-item version for all further analyses.

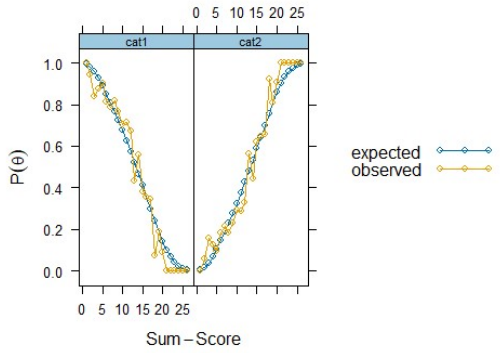
Table 3

Item Fit Final Hierarchical IRT Model

Item	S_ χ^2 Statistic	df	RMSEA	FDR adjusted p-value
VR4	28.1	12	.041	.048
VR5	35.0	14	.043	.020
NS4	35.9	10	.056	.002

Figure 1

Item Fit Plot Item VR4: Observed versus Expected Values



[10] We use the cutoff values of Hu and Bentler (1999) to evaluate the goodness of fit: RMSEA <.06, SRMR <.08, and CFI & TLI > .95 indicate good fit.

Figure 2

Item Fit Plot Item VR5: Observed versus Expected Values

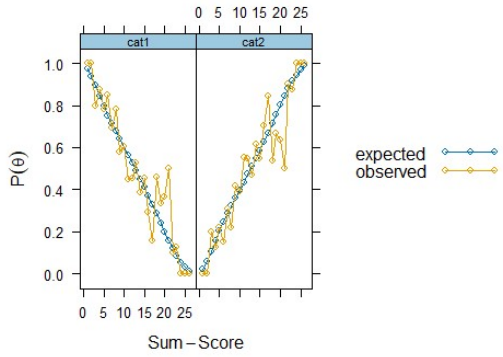


Figure 3

Item Fit Plot Item NS4: Observed versus Expected Values

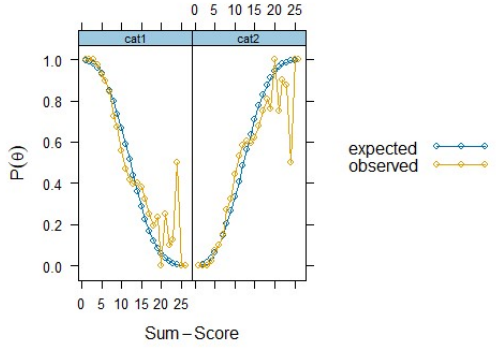


Table 4

Standardized and Unstandardized Slopes Final Hierarchical IRT Model

Item	Unstd. Intercept	Unstd. MR slope	Std. MR slope	Unstd. VR slope	Std. VR slope	Unstd. NS slope	Std. NS slope	Unstd. FA slope	Std. FA slope
MR1	-0.530	0.465	0.231	0.000	0.000	0.000	0.000	0.000	0.000
MR2	-0.030	1.120	0.444	0.000	0.000	0.000	0.000	0.000	0.000
MR3	-1.316	0.595	0.285	0.000	0.000	0.000	0.000	0.000	0.000
MR4	-1.082	0.849	0.373	0.000	0.000	0.000	0.000	0.000	0.000
MR6	-1.036	0.618	0.294	0.000	0.000	0.000	0.000	0.000	0.000
MR7	-1.726	0.456	0.227	0.000	0.000	0.000	0.000	0.000	0.000
VR1	1.742	0.000	0.000	0.589	0.229	0.000	0.000	0.000	0.000

Item	Unstd. Intercept	Unstd. MR slope	Std. MR slope	Unstd. VR slope	Std. VR slope	Unstd. NS slope	Std. NS slope	Unstd. FA slope	Std. FA slope
VR2	1.312	0.000	0.000	1.009	0.344	0.000	0.000	0.000	0.000
VR3	-0.075	0.000	0.000	0.956	0.332	0.000	0.000	0.000	0.000
VR4	-0.724	0.000	0.000	1.392	0.414	0.000	0.000	0.000	0.000
VR5	-0.362	0.000	0.000	0.838	0.303	0.000	0.000	0.000	0.000
VR6	-1.578	0.000	0.000	0.665	0.253	0.000	0.000	0.000	0.000
VR7	-1.532	0.000	0.000	0.532	0.210	0.000	0.000	0.000	0.000
VR8	-1.531	0.000	0.000	1.058	0.355	0.000	0.000	0.000	0.000
NS1	1.827	0.000	0.000	0.000	0.000	0.884	0.242	0.000	0.000
NS2	0.058	0.000	0.000	0.000	0.000	1.106	0.283	0.000	0.000
NS3	0.349	0.000	0.000	0.000	0.000	1.729	0.360	0.000	0.000
NS4	-0.671	0.000	0.000	0.000	0.000	2.274	0.397	0.000	0.000
NS5	-0.428	0.000	0.000	0.000	0.000	3.067	0.428	0.000	0.000
NS7	-1.572	0.000	0.000	0.000	0.000	1.331	0.316	0.000	0.000
NS8	-2.067	0.000	0.000	0.000	0.000	1.214	0.300	0.000	0.000
FA1	-0.370	0.000	0.000	0.000	0.000	0.000	0.000	1.651	0.729
FA2	-0.075	0.000	0.000	0.000	0.000	0.000	0.000	1.919	0.753
FA3	-0.164	0.000	0.000	0.000	0.000	0.000	0.000	2.344	0.778
FA4	0.174	0.000	0.000	0.000	0.000	0.000	0.000	1.218	0.665
FA5	-2.665	0.000	0.000	0.000	0.000	0.000	0.000	0.378	0.315
FA6	-1.756	0.000	0.000	0.000	0.000	0.000	0.000	0.626	0.467

Note. VR = Verbal Reasoning, MR = Matrix Reasoning, NS = Number Series, FA = Figural Analogies.

Table 5

Item-Wise Score-Based DIF Detection Final Hierarchical IRT Model

item	item type	number of est. parameters	LM test statistic	p-value	FDR adjusted p-value
MR1	2PL	2	1.962	0.375	0.562
MR2	2PL	2	2.401	0.301	0.562
MR3	2PL	2	2.256	0.324	0.562
MR4	2PL	2	1.268	0.531	0.677
MR6	2PL	2	5.239	0.073	0.328
MR7	2PL	2	0.379	0.827	0.894
VR1	2PL	2	0.563	0.755	0.849
VR2	2PL	2	2.771	0.250	0.562
VR3	2PL	2	0.235	0.889	0.896

item	item type	number of est. parameters	LM test statistic	p-value	FDR adjusted p-value
VR4	2PL	2	7.652	0.022	0.147
VR5	2PL	2	1.101	0.577	0.677
VR6	2PL	2	3.157	0.206	0.506
VR7	2PL	2	2.031	0.362	0.562
VR8	2PL	2	10.624	0.005	0.069
NS1	2PL	2	3.911	0.142	0.382
NS2	2PL	2	5.808	0.055	0.296
NS3	2PL	2	4.216	0.121	0.382
NS4	2PL	2	10.047	0.007	0.069
NS5	2PL	2	2.090	0.352	0.562
NS7	2PL	2	9.731	0.008	0.069
NS8	2PL	2	4.370	0.112	0.382
FA1	2PL	2	1.515	0.469	0.666
FA2	2PL	2	4.027	0.134	0.382
FA3	2PL	2	0.219	0.896	0.896
FA4	2PL	2	1.105	0.575	0.677
FA5	2PL	2	1.124	0.570	0.677
FA6	2PL	2	2.144	0.342	0.562

Internal consistency

The full Ch-ICAR showed good internal consistency ($\alpha = .82$; $\omega = .82$, 95% CI_ω [.81, .84]). At subtest level, two of the four subtests also showed good internal consistency: Number Series ($\alpha = .73$; $\omega = .75$, 95% CI_ω [.72, .77]) and Figural Analogies ($\alpha = .77$; $\omega = .81$, 95% CI_ω [.79, .83]). The internal consistency of the other two subtests was less favorable: Verbal Reasoning had $\alpha = .60$; $\omega = .61$, 95% CI_ω [.57, .65] and Matrix Reasoning had $\alpha = .48$; $\omega = .48$, 95% CI_ω [.42, .54].

Distribution of Ch-ICAR scores

Table 6 presents the means and standard deviations of the Ch-ICAR (subtest) scores, with separate values provided for boys and girls. Notably, the maximum scores are 8 for Verbal Reasoning, 6 for both Matrix Reasoning and Figural Analogies, 7 for Number Series, and 27 for the full Ch-ICAR. Results of the independent samples t-test showed no significant

difference in Ch-ICAR total score between males (M = 10.91, SD = 5.20) and females (M = 10.54, SD = 4.99): $t(812) = 1.02$, $p = .31$, *Cohen's d* = .07. The histogram of the standardized Ch-ICAR total scores is depicted in Figure 4.

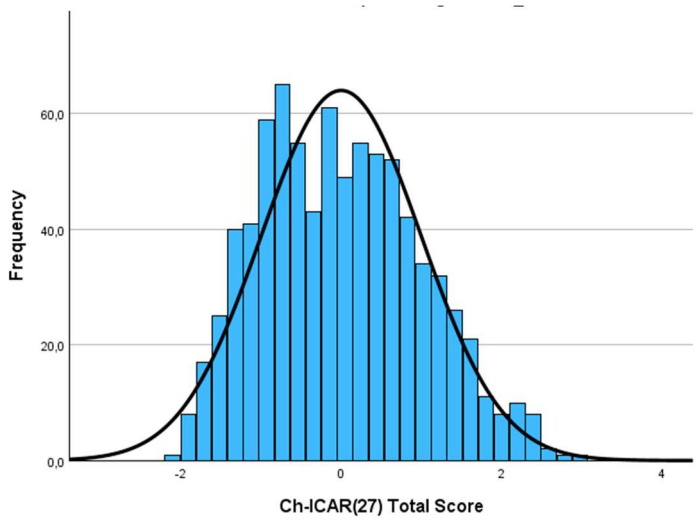
Table 6

Means and Standard Deviations of the Ch-ICAR (Subtest) Scores

(sub)test	M (*SD*)	M *(SD)* Boys	M *(SD)* Girls
Verbal Reasoning	3.49 *(1.81)*	3.64 *(1.86)*	3.28 *(1.74)*
Matrix Reasoning	1.86 *(1.42)*	1.79 *(1.45)*	1.94 *(1.39)*
Number Series	3.19 *(1.99)*	3.41 *(2.00)*	2.93 *(1.95)*
Figural Analogies	2.22 *(1.85)*	2.07 *(1.81)*	2.40 *(1.88)*
Full Ch-ICAR	10.76 *(5.11)*	10.91 *(5.20)*	10.54 *(4.99)*

Figure 4

Histogram of Standardized Ch-ICAR Total Scores



Relation with academic performance

All measures of academic performance were positively and significantly correlated with both the subtests and the full Ch-ICAR (Table 7). Correlations with CDR were

consistently stronger than those with the other measures of academic performance. However, given that two out of four subtests demonstrate limited reliability, we recommend caution when interpreting individual subtest scores independently.

Table 7

Correlations Between Academic Performance and the Ch-ICAR (Sub)tests

Academic performance	Verbal Reasoning	Matrix Reasoning	Number Series	Figural Analogies	Full Ch-ICAR
GPA	.35**	.27**	.29**	.28**	.42**
Dutch	.33**	.23**	.28**	.25**	.38**
Mathematics	.40**	.29**	.38**	.32**	.49**
Natural sciences	.34**	.23**	.30**	.25**	.40**
Technics	.20**	.21**	.14**	.20**	.26**
CDR	.58**	.31**	.54**	.34**	.63**

Note. ** $p < .01$.

Subsequently, we conducted three mixed effects models, each with a different dependent variable: GPA, mathematics score and CDR score. In each model, we included the Ch-ICAR total score (continuous), sex (categorical) and all four SES indicators[11] (all categorical) as fixed effects. Additionally, school was included as a random factor in each model to account for the multilevel nature of the data. To address potential multicollinearity concerns, we assessed the variance inflation factor (VIF) for all predictors. All VIF values ranged from 1.01 to 1.30, which is well below the commonly cited thresholds[12] for problematic multicollinearity in the model (Marcoulides & Raykov, 2019; O'Brien, 2007). Table 8 includes the results of the mixed effects models. The models achieved a marginal R^2 of .32, .34 and .40 for GPA, mathematics score and CDR score as dependent variable

[11] All SES indicators were significantly correlated with GPA, math score and CDR score, with the exception of *Educational Level Parent 1*, which showed no correlation with GPA and math score. See Appendix E for the full correlation table.

[12] "Not uncommonly a VIF of 10 or even one as low as 4 have been used as rules of thumb to indicate excessive or serious multicollinearity" (O'Brien, 2007, p.674).

respectively, indicating that up to 40% of the variance in academic performance is explained by the fixed effects portion of the model (Nakagawa & Schielzeth, 2013; Shaw et al., 2023). The Ch-ICAR total score uniquely explained 18%, 26% and 35% of the within school variance in GPA, math score and CDR score respectively via its fixed effects.

Table 8

Estimates, Standard Errors, t-values and p-values for the Fixed and Random Effects of the Three Linear Mixed Effect Models

	GPA				Mathematics Score				CDR Score			
	β	SE	*t*	*p*	β	SE	*t*	*p*	β	SE	*t*	*p*
Fixed Effects												
Intercept	-.99	.11	-8.94	<.001	-1.07	.11	-9.42	<.001	-1.27	.09	-13.73	<.001
Ch-ICAR	.08	.01	11.94	<.001	.10	.01	14.40	<.001	.12	.01	18.38	<.001
Sex	.50	.07	7.03	<.001	.14	.07	1.98	.05	.18	.06	2.83	<.01
Language	-.06	.14	-.40	.69	-.13	.14	-.90	.37	-.24	.13	-1.94	.05
Bursary	-.47	.08	-5.68	<.001	-.45	.08	-5.39	<.001	-.06	.07	-.80	.42
Education Parent 1	.18	.14	1.25	.21	.22	.14	1.56	.12	-.10	.12	-.84	.40
Education Parent 2	-.31	.15	-2.08	.04	-.15	.15	-1.04	.30	-.24	.12	-2.03	.04
Random Effects	Variance	SD			Variance	SD			Variance	SD		
School												
(Intercept)	.04	.19			.05	.21			.01	.10		

Note. Reference level for Sex = boys; reference level for Language = Dutch as home language; reference level for Bursary = no bursary; reference level for Education parent 1 and 2 = highest attained degree is secondary education or higher.

Discussion

In Study 1, we aimed to evaluate the psychometric properties of the Ch-ICAR by examining four key aspects of the test. Results from the final hierarchical IRT analysis, excluding four items (MR5, MR8, NS6, and FA7), provided support for the Ch-ICAR's psychometric robustness. All remaining items demonstrated adequate discriminatory power, and no DIF was detected. Furthermore, the good model fit aligns with expectations based on CHC theory, which posits several broad cognitive abilities at the first, lower level and general cognitive ability at the second, higher level (Flanagan & McDonough, 2018). We thus utilized the 27-item subset in all subsequent analyses and recommend that future research also adopt this subset when using the Ch-ICAR.

Regarding internal consistency, analyses revealed that the full Ch-ICAR shows good internal consistency, which contributes to the quality of the test. At the subtest level, however, reliability estimates are mixed: While Number Series and Figural Analogies demonstrate good internal consistency, Matrix Reasoning and Verbal Reasoning exhibit insufficient and marginally sufficient reliability respectively. These findings align with prior research on the original ICAR16, where similar[13] reliability challenges were reported for Matrix Reasoning ($\omega = .55$) and Verbal Reasoning ($\omega = .61$; Condon & Revelle, 2014; Young et al., 2019). The consistent pattern of low reliabilities across studies suggests that these two subtests may be capturing multiple underlying factors rather than a single dimension. Given that two out of four subtests demonstrate limited reliability, we recommend caution when interpreting individual subtest scores independently. However, reliability and validity are always a function of both the instrument and the specific population being tested (Bannigan & Watson, 2009), so future research is essential to assess whether these subtests might perform more

[13] Of course, reliability is influenced by the length of a test, so comparisons between tests of differing lengths should be made with careful consideration and nuance.

reliably in other populations.

With respect to the distribution of the Ch-ICAR scores, our analysis showed no evidence of sex bias. This result aligns with previous research, which suggests that although small sex differences in specific cognitive skills may exist, overall cognitive ability is generally equal across sexes (e.g., Deary et al., 2007; Fergusson & Horwood, 1997; Nisbett et al., 2012).

Regarding the relation with academic performance, we found that cognitive ability as measured by the Ch-ICAR uniquely explained about 18% of the within school variance in GPA, 26% of the within school variance in the score for the school subject mathematics, and 35% of the within school variance in the score on the CDR. The proportion of explained variance in school achievement is somewhat lower than what is commonly reported for educational tests (i.e. medium to high, depending on the way academic performance is measured and analyzed, see for instance Kort et al., 2002; Deary et al., 2007; Koenig et al., 2008), but acceptable considering the rather short nature of the Ch-ICAR. Moreover, it is crucial to highlight that we have controlled for SES and sex, and accounted for the nested structure of our data, enhancing the robustness of the results. Furthermore, the correlations in our study are similar to the magnitude of the correlations with academic achievement found in the initial validation study of the ICAR16 (Condon & Revelle, 2014), which confirms the present set as a useful contribution to the ICAR project.

In conclusion, the Ch-ICAR seems promising as a quick, cost-efficient (i.e., free to use) and useful tool for researchers to obtain a reliable and valid, sex-neutral assessment of cognitive ability.

Study 2: Cross-validation of the Children's ICAR

Study 2 sought to cross-validate the Ch-ICAR with two established intelligence tests: the Raven's 2 Progressive Matrices (RPM; McLeod & McCrimmon, 2021) and the CoVaT-CHC Basic Version (CoVaT; Magez et al., 2015). We examined the relation between the Ch-ICAR and cognitive ability, estimated by RPM. Additionally, we examined the relations between the Ch-ICAR subtests and three broad cognitive abilities estimated by the CoVaT: Fluid intelligence (Gf), Visuospatial processing (Gv), and Crystallized intelligence (Gc), as well as the relation between the Ch-ICAR and the CoVaT total score. A priori we expected significant correlations of the Ch-ICAR with on the one hand RPM and on the other CoVaT.

As mentioned in the general introduction, the ICAR project does not start from an a priori theoretical framework of cognitive ability. Only one study explored the connection between the original four ICAR-(sub)tests and the CHC-model (Young & Keith, 2020). Results showed a strong correlation between the ICAR16 total score and general cognitive ability, estimated by the WAIS-IV. At subtest level, Young and Keith (2020) found that Letter and Number Series correlated most strongly with tests of Gf, whereas the remaining three subtests (Verbal Reasoning, Three-dimensional Rotation, and Matrix Reasoning) showed the strongest correlations with tests of Gv. Regarding Verbal Reasoning, Young and Keith (2020) themselves expressed surprise at the results given the inconsistency with the CHC-theory. They considered it highly plausible that the results stemmed from a statistical artifact or the small sample size, and hence pleaded for replications. Regarding Matrix Reasoning, the authors suggested that it is likely that it calls on both Gf and Gv but recommended caution in interpretation given the low internal consistency.

Building on cognitive ability theory, we anticipated that the Ch-ICAR subtests Number Series and Matrix Reasoning mainly tap into Gf; that Figural Analogies draws on both Gf and Gv; and that Verbal Reasoning calls on Gf and Gc, since this subtest requires

both mathematical operations and general knowledge. However, we did not rule out the possibility that the results regarding Verbal Reasoning and Matrix Reasoning are in line with the earlier research of Young and Keith (2020), and primarily assess Gv. It is important to note here that the Ch-ICAR does not aspire to measure IQ like the Wechsler scales. Instead, the Ch-ICAR strives to be a concise, open-source measure of cognitive ability for research purposes, in line with the overarching aim of the ICAR project (Condon & Revelle, 2014).

Method

Participants

Data for cross-validation with RPM were collected from 91 pupils (N_{RPM} = 91, male 40%), who were in their second year of secondary education (M_{age} = 12.99, SD_{age} = .57). In this sample, 3% of the participants had another home language than Dutch. Data for cross-validation with the CoVaT were collected from 96 pupils (N_{CoVaT} = 96, male 50%), who were in their last year of primary education (M_{age} = 11.05, SD_{age} = .27). In this sample, 3% of the participants had another home language than Dutch. Table 2 offers a comprehensive overview of the demographics and associated statistics for each sample.

Materials

Children's ICAR. All pupils completed the Ch-ICAR (see Study 1 for a detailed description of the instrument). Based on recommendations from Study 1, we utilized the 27-item subset for all analyses. The means and standard deviations of the Ch-ICAR (subtest) scores for the CoVaT and RPM samples are presented in Tables 9 and 10 respectively, with separate values reported for boys and girls[14]. The full Ch-ICAR showed good internal

[14]In Flanders, primary education is comprehensive, with no ability grouping and open access, while secondary education is specialized and ability-based (Seghers et al., 2019). Study 2 focused on examining the Ch-ICAR's correlation with other cognitive ability tests, rather than sampling a representative population. Consequently, sex balance across ability groups in the RPM sample was not specifically monitored, as this was not a focus of the study. Nevertheless, independent samples t-tests showed no significant difference in Ch-ICAR total scores between males and females in either sample: CoVaT sample: $t(94) = .82$, $p = .41$, *Cohen's d* = .17; RPM sample: $t(89) = -1.48$, $p = .14$, *Cohen's d* = -.32. Furthermore, no DIF was detected in any items in the refined hierarchical IRT model (excluding items MR5, MR8, NS6 and FA7).

consistency in both samples, with McDonald's omega values of .81 in the RPM sample and .84 in the CoVaT sample. At subtest level, internal consistency was sufficient to good for three of the four subtests (Number Series, Figural Analogies and Verbal Reasoning) in both samples, with McDonald's omega values ranging from .60 to .77. However, internal consistency for the Matrix Reasoning subtest was insufficient (ω = .31 in the RPM sample; ω = .52 in the CoVaT sample). Comprehensive internal consistency statistics, including both McDonald's omega and Cronbach's alpha, are provided in Appendix F.

Table 9

Means and Standard Deviations of the Ch-ICAR (Subtest) Scores for CoVaT Sample

(sub)test	M (*SD*)	M (*SD*) Boys	M *(SD)* Girls
Verbal Reasoning	3.93 *(1.78)*	4.17 *(2.01)*	3.69 *(1.49)*
Matrix Reasoning	2.55 *(1.52)*	2.46 *(1.52)*	2.65 *(1.54)*
Number Series	3.58 *(1.98)*	4.13 *(1.94)*	3.04 *(1.88)*
Figural Analogies	2.22 *(1.80)*	1.98 *(1.76)*	2.46 *(1.82)*
Full Ch-ICAR	12.28 *(5.32)*	12.73 *(5.53)*	11.83 *(5.12)*

Note. The maximum scores are 8 for Verbal Reasoning, 6 for both Matrix Reasoning and Figural Analogies, 7 for Number Series, and 27 for the full Ch-ICAR.

Table 10

Means and Standard Deviations of the Ch-ICAR (Subtest) Scores for RPM Sample

(sub)test	M (*SD*)	M (*SD*) Boys	M *(SD)* Girls
Verbal Reasoning	3.98 *(1.99)*	3.61 *(1.78)*	4.22 *(2.10)*
Matrix Reasoning	2.51 *(1.34)*	2.42 *(1.50)*	2.56 *(1.23)*
Number Series	4.12 *(2.06)*	4.11 *(2.05)*	4.13 *(2.08)*
Figural Analogies	2.98 *(1.71)*	2.47 *(1.87)*	3.31 *(1.53)*
Full Ch-ICAR	13.58 *(5.09)*	12.61 *(5.40)*	14.22 *(4.83)*

Note. The maximum scores are 8 for Verbal Reasoning, 6 for both Matrix Reasoning and Figural Analogies, 7 for Number Series, and 27 for the full Ch-ICAR.

Raven's 2 Progressive Matrices. We used the Raven's 2 Digital Short Form (McLeod & McCrimmon, 2021). The test starts with three example items followed by three practice items. The actual test consists of 24 items and has a time limit of 20 minutes. Each participant received a unique but comparable subset of items, automatically generated from the Raven's 2 item bank. The marginal reliability coefficient of the Raven's 2 Digital Short Form, based on American data, is .80 (Dimitrov, 2003). Per participant, a score report was obtained via Q-Global (Pearson, 2023). The report describes among others: a Total Raw Score (TRS), a Skill Score (SkS) and a Scaled Score (ScS). For our analyses we used the ScS[15]. In our sample, the mean ScS was $M = 96.44$, $SD = 12.18$.

CoVaT-CHC Basic Version. The full CoVaT (Magez et al., 2015) encompasses five broad cognitive abilities (BCA): Fluid intelligence (Gf), Visuospatial processing (Gv), Crystallized intelligence (Gc), Short-Term Memory (Gsm), and Processing Speed (Gs). Each BCA is measured by two subtests, except Gs, which is measured by only one. The CoVaT offers strong psychometric properties (Magez, 2019) and has received the highest quality label of the Belgian Federation for Psychologists (BFP, 2020). The complete CoVaT administration takes two and a half to three hours, but individual administration may be shorter. Due to time constraints, we only administered the six subtests measuring Gf, Gv, and Gc.

Gf is assessed by the subtests Point Sequences and Figure Sequences. Point Sequences comprises 15 items where participants have to draw points to complete a sequence. Figure Sequences comprises 25 items and requires participants to draw figures to complete a sequence. Gc is measured via the subtests Shifts and Opposites, each consisting of 35 items.

[15] The Skill Score converts each participant's test performance to a common and equal interval scale, regardless of test version or differences between item sets (Pearson, 2023). Skill Scores can be directly compared with each other. The Scaled Score is based on the Skill Score and the participant's age. The Scaled Score is a standardized score with a mean of 100 and a standard deviation of 15, and can be compared to the score on other intelligence tests.

In Shifts, participants have to select the word that does not fit in a series, while Opposites requires participants to select the opposite word from a target word. Gv is evaluated using the subtests Rotated Figures (20 items) and Folding Boxes (26 items), assessing participants' ability to identify equal, albeit rotated, figures and match folded with unfolded boxes, respectively. Each subtest has a specified time limit outlined in the manual. The reliability coefficients of Gf, Gv, and Gc are .96, .91 and .87 respectively[16]. The means and standard deviations of the BCAs in our sample are shown in Table 11. We used the pen-and-paper version and administered the instrument in group.

Table 11

Means and Standard Deviations of the CoVaT (BCAs)

CoVaT (BCA)	*M* (*SD*)
Gv	54.76 *(17.59)*
Gf	67.32 *(14.97)*
Gc	54.42 *(9.86)*
CoVaT weighted overall score[17]	298.24 *(54.84)*

Procedure

For cross-validation with RPM, participants completed RPM and the Ch-ICAR on the same day. For cross-validation with the CoVaT, participants completed the CoVaT and the Ch-ICAR on different days, with minimum 7 and maximum 35 days between both tests (M_{days} = 16.33, SD_{days} = 16.17). We introduced a time delay specifically for the Ch-ICAR-CoVaT test-taking to prevent pupils from experiencing cognitive overload before beginning the

[16] The reliability of the BCA indices was estimated based on the formula of Lienert (Stinissen et al., 1975), which uses the observed standard deviations (σ) per subtest and the observed intercorrelations between subtests.
[17] The weighted CoVaT overall score = 2*Gf + 2*Gc + Gv.

subsequent test, given that the CoVaT spanned about three hours. Since the Raven only lasted 20 minutes, no time delay was deemed necessary between the Raven and the Ch-ICAR. The order of testing alternated between classes, in both samples.

Analyses

Analyses were conducted in SPSS (version 29.0; IMB Corp., 2022) and R (version 4.3.3; R Core Team, 2024). All R analysis codes as well as the variance-covariance matrix are available at: https://osf.io/6cwfs/?view_only=4a80feb83bd94c378d3bd06853820563.

To examine the relationship between the Ch-ICAR and cognitive ability, we calculated Pearson correlations between the observed Ch-ICAR (subtest) scores and the Scaled Score derived from the RPM, as well as between the observed Ch-ICAR (subtest) scores and the CoVaT's weighted overall score, and its measures of the broad cognitive abilities Gf, Gv, and Gc. In line with Young and Keith's (2020) cross-validation research, we also tested a final integrated CFA model informed by the theoretical expected relationships. The CFA was conducted using the lavaan package (version 0.6.17; Rosseel, 2012), with maximum likelihood (ML) as estimator and the Satorra-Bentler correction for non-normality. The model allowed for correlations between the broad cognitive abilities to account for the positive manifold (i.e., the psychometric phenomenon that all (sub)tests that measure facets of cognitive ability correlate; Burgoyne et al., 2022), and for correlated errors for the Ch-ICAR subtests to account for potential method covariance (Young & Keith, 2020).

Results

Cross-validation with the RPM

The correlations between the Ch-ICAR subtests and the ScS range from r = .39 to r = .59 (Table 12). The Ch-ICAR total score showed the largest correlation with the ScS: r = .67.

Table 12

Correlations Between the Ch-ICAR and RPM

(sub)test	1	2	3	4	5	6
1. RPM: ScS	1					
2. Ch-ICAR: VR	.59**	1				
3. Ch-ICAR: MR	.43**	.46**	1			
4. Ch-ICAR: NS	.48**	.47**	.33**	1		
5. Ch-ICAR: FA	.39**	.30**	.34**	.19	1	
6. Full Ch-ICAR	.67**	.80**	.69**	.74**	.62**	1

Note. ScS = Scaled Score, VR = Verbal Reasoning, MR = Matrix Reasoning, NS = Number Series, FA = Figural Analogies.
**$p < .01$.

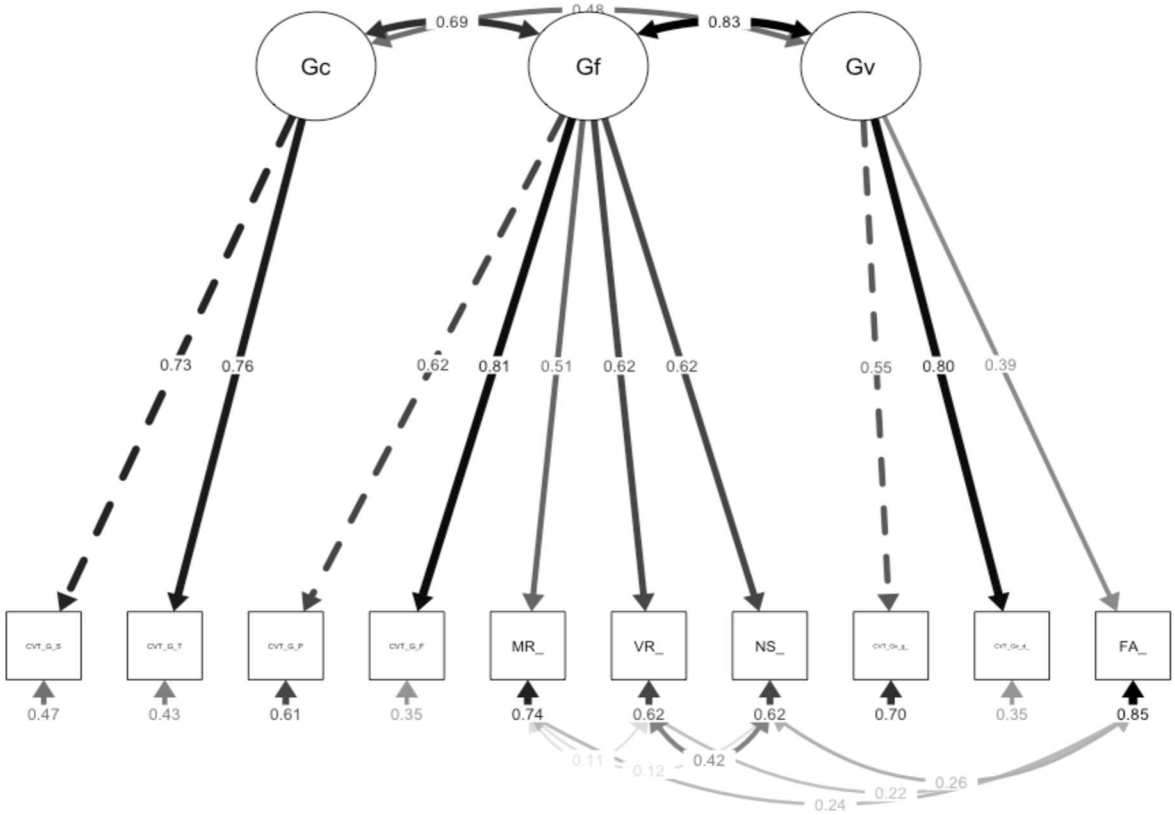
Cross-validation with the CoVaT

The correlations between the Ch-ICAR and CoVaT scores are depicted in Table 13. Number Series, Verbal Reasoning and Matrix Reasoning showed the strongest correlation with Gf, while Figural Analogies correlated most strongly with Gv, when considering all BCAs. However, given that two out of four subtests demonstrate limited reliability, we recommend caution when interpreting individual subtest scores independently.

The final integrated CFA model (Figure 5) showed a moderate fit for the data: χ^2 (26): 41.305; RMSEA = .081, 90% C.I. [.026, .125]; CFI = .944, TLI = .904, SRMR = .060. Given that we a priori expected Verbal Reasoning and Figural Analogies to show equally high correlations with Gf and Gc; and Gf and Gv respectively, we initially allowed those two subtests to crossload on both. However, compared to the final integrated model, crossloading Verbal Reasoning did not significantly improve the model fit: $\chi^2(25)$: 39.574; RMSEA = .081, 90% C.I. [.024, .126]; CFI = .947, TLI = .904, SRMR = .058; $\Delta\chi^2(1) = 1.774$, $p = .183$. Similarly, crossloading Figural Analogies also had no significant impact: χ^2 (25): 40.667; RMSEA = .084, 90% C.I. [.030, .129]; CFI = .942, TLI = .896, SRMR = .061, $\Delta\chi^2(1) = .085$, $p = .771$. Therefore, we opted for the more parsimonious model without crossloadings.

Figure 5

Final Integrated CFA Model



Note. CVT_... = CoVaT subtest, VR_ = Verbal Reasoning scale score, MR_ = Matrix Reasoning scale score, NS_ = Number Series scale score, FA_ = Figural Analogies scale score.

Table 13

Correlations Between the Ch-ICAR (Sub)tests, the CoVaT Overall Score and Three Broad Cognitive Abilities Estimated by the CoVaT

(sub)test	1	2	3	4	5	6	7	8	9
1. CoVaT weighted overall score	1								
2. CoVaT: Gv	.79**	1							
3. CoVaT: Gf	.90**	.63**	1						
4. CoVaT: Gc	.71**	.35**	.43**	1					
5. Ch-ICAR: VR	.54**	.28**	.53**	.45**	1				
6. Ch-ICAR: MR	.47**	.38**	.40**	.37**	.40**	1			
7. Ch-ICAR: NS	.55**	.35**	.51**	.44**	.62**	.39**	1		
8. Ch-ICAR: FA	.31**	.29**	.26*	.21*	.34**	.35**	.40**	1	
9. Full Ch-ICAR	.62**	.43**	.57**	.49**	.79**	.68**	.82**	.70**	1

Note. VR = Verbal Reasoning, MR = Matrix Reasoning, NS = Number Series, FA = Figural Analogies.
*$p < .05$, **$p < .01$.

Discussion

In Study 2, we cross-validated the Ch-ICAR with the RPM in a sample of 91 pupils in secondary education; and with the CoVaT in a sample of 96 pupils in primary education. Analyses revealed high correlations between the Ch-ICAR total score and cognitive ability, estimated by RPM and CoVaT. The magnitude of the relationships is in line with prior cross-validation research with the ICAR16 (Young & Keith, 2020). As expected, the Ch-ICAR is mainly a measure of Gf, making it a viable measure for non-verbal cognitive ability, specifically.

At subtest level, the correlation matrix in combination with the integrated CFA model suggests that the subtests tend to measure what we theoretically expected. We found Verbal Reasoning to predominantly assess Gf and, to a somewhat lesser extent, Gc. These results are consistent with the theory and empower the statement of Young and Keith (2020) that their findings (Verbal Reasoning mainly tapping into Gv) are presumably a statistical artifact or a product of the small sample.

Regarding Figural Analogies, it is notable that this subtest shows rather small correlations with the CoVaT across the board. When we take a closer look, several reasons emerge as to why this subtest does not function as expected. First, the scores on this subtest are not normally distributed at all (Appendix G); many participants either get four questions right or none at all. Secondly, and related to the first reason, there appears to be an issue with increasing difficulty: While about half of the participants answered items 1-4 correctly, the correct response rate dropped to 11% for item 5 and 20% for item 6 (Appendix H displays the proportion of correct responses for each item across all samples). Additionally, as we aimed for standardization in the testing procedure, we presented the subtests in a fixed sequence, with Figural Analogies as the final subtest. While this procedure adheres to common practice (Flanagan & Alfonso, 2017; Weiss et al., 2019), it introduces the potential for fatigue effects

to impact test performance. To address these issues, we recommend conducting further research to assess the validity of Figural Analogies with other items of the subtest that vary in difficulty.

The subtest Matrix Reasoning correlates most strongly with Gf, as expected; however, its low reliability requires careful interpretation. Future research should explore ways to enhance the quality of Matrix Reasoning, as its validity issues reflect broader challenges within the ICAR project (Condon & Revelle, 2014, p. 57; Young et al., 2019, p. 2). For the Ch-ICAR specifically, developing a larger set of easier items and incorporating simple, non-verbal instructions may enhance the subtest's internal consistency and further improve the overall quality of the Ch-ICAR. Additionally, one could consider replacing Matrix Reasoning items with alternative item types, such as Progressive Matrices items, as those use a similar task paradigm but offer much stronger reliability.

In conclusion, the Ch-ICAR total score exhibits a strong correlation with cognitive ability, establishing itself as a valid and freely available measure for assessing children's cognitive abilities in research settings. Nonetheless, at this stage, we advise against using and interpreting subtests independently.

As the sample sizes of Study 2 are rather small, the current results should be interpreted with caution and replications in larger samples are needed. Nevertheless, the results of both cross-validations yield the same conclusions, which removes part of the potential bias risk.

General conclusion

In developing the Ch-ICAR, our aim was to expand the scope of the ICAR project, considering that its suitability for children had not been previously assessed (Dworak et al., 2021; The International Cognitive Ability Resource Team, 2014). Additionally, we sought to

address the dearth of psychometrically robust cognitive ability measures that are freely accessible, brief in administration time, and tailored to children.

The results of Study 1 and 2 provide support for the utility of the full Ch-ICAR as a measurement for children's cognitive abilities within a research context. This was demonstrated by the observed correlations between the Ch-ICAR and the RPM and CoVaT, two established intelligence tests (McLeod & McCrimmon, 2021; Magez et al., 2015) and by the established relationship with academic performance. The magnitude of the correlation coefficients is in line with prior validation research of the ICAR16 (Condon & Revelle, 2014; Young & Keith, 2020).

At this point, however, we advise against the use of the Ch-ICAR subtests in isolation, in contrast with the use of the battery as a whole. Further research should ascertain whether the subtests can reliably stand alone and yield meaningful interpretations. Therefore, replications of the relations with the broad cognitive abilities are necessary, as well as improvements in the subtests Matrix Reasoning and Figural Analogies.

The primary constraint of this investigation pertains to the restricted scope of the Ch-ICAR, which solely assessed a limited subset of ICAR items in children. Consequently, the possibility cannot be dismissed that an alternative or more extensive item selection may yield superior performance as a child-adapted version of the ICAR. Another consideration involves the absence of established norms for the Ch-ICAR. Given the dynamic developmental changes in cognitive processing efficiency during childhood and adolescence (Flanagan & McDonough, 2018), it is imperative to consider factors such as age or grade level when interpreting and comparing Ch-ICAR results. Therefore, we underscore the significance of future research endeavors aimed at establishing age- or grade-based norms to ensure precise interpretation and comparison of Ch-ICAR outcomes.

Declarations

Funding

This work was funded by Voka Leerstoel Studie-oriëntering, WBS-element = E/01447/52

Acknowledgments

We would like to thank Sari Van Haute and Eline Rombaut for their contribution to data-collection. We also extend our sincere gratitude to the reviewers for their valuable comments and suggestions, which have significantly enriched our analyses and discussion.

Conflicts of interest/Competing interests

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

Ethics approval

This research was performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of the Faculty of Psychology and Educational Sciences of Ghent University (2 September 2021/ No 2021/59).

Consent to participate

Written informed consent was obtained from the parents.

Consent for publication

Not applicable

Availability of data and materials

For now (pre-peer review), the Ch-ICAR can be accessed via this link: https://ap.lc/UXMoM, with this access data: Username: PeerReviewChICAR; Password: 2HE2wiA1. Upon acceptance, the Ch-ICAR will be available via the homepage, users will only need to create a user account. The variance-covariance matrices are available at https://osf.io/6cwfs/?view_only=4a80feb83bd94c378d3bd06853820563; the complete dataset is available upon request; none of the studies were preregistered.

Code availability

Analysis codes are available at https://osf.io/6cwfs/?view_only=4a80feb83bd94c378d3bd06853820563

Authors' contributions

The authors confirm contribution to the paper as follows. Study conception and design: Derous Eva, Dirix Nicolas, Dutry Merel, Duyck Wouter, Schelfhout Stijn, Schittekatte Mark, Vereeck Alexandra. Data collection: Merel Dutry. Analysis and interpretation of results: Derous Eva, Dirix Nicolas, Dutry Merel, Duyck Wouter, Schelfhout Stijn, Schittekatte Mark, Szmalec Arnaud, Vereeck Alexandra, Woumans Evy, Dries Debeer. Draft manuscript preparation: Dutry Merel, Vereeck Alexandra. All authors reviewed the results and approved the final version of the manuscript.

References

Ali, A., Ambler, G., Strydom, A., Rai, D., Cooper, C., McManus, S., … Hassiotis, A. (2013). The relationship between happiness and intelligent quotient: The contribution of socio-economic and clinical factors. *Psychological Medicine, 43,* 1303-1312. https://doi.org/10.1017/S0033291712002139

Aubry, A., & Bourdin, B. (2018). Short forms of Wechsler scales assessing the intellectually gifted children using simulation data. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00830

Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing, 18,* 3237-3243. https://doi.org/ 10.1111/j.1365-2702.2009.02939.x

Bartoń, K. (2023). *MuMIn: Multi-Model Inference* (Version 1.47.5) [Software]. Retrieved from https://cran.r-project.org/web/packages/MuMIn/index.html

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 57*, 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

BFP Testcommissie (2020). Testfiche CoVaT-CHC Basisversie. BFP. https://www.bfp-fbp.be/_files/ugd/8d37d9_8647ceeb137449b5b967fbf48fa89f9e.pdf

Burgoyne, A. P., Mashburn, C. A., Tsukahara, J. S., & Engle, R. W. (2022). Attention control and process overlap theory: Searching for cognitive processes underpinning the positive manifold. *Intelligence, 91,* 101629. https://doi.org/10.1016/j.intell.2022.101629

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*, 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2023). *mirt: Multidimensional item response theory* (Version 1.35) [R package]. https://cran.r-project.org/web/packages/mirt/mirt.pdf

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

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*, 13–21. https://doi.org/10.1016/j.intell.2006.02.001

Debeer, D. (2020). *scDIFtest: Item-Wise Score-Based DIF Detection* (Version 0.1.1) [R package]. https://cran.r-project.org/web/packages/scDIFtest/scDIFtest.pdf

Desoete, A., & Roeyers, H. (2006). *Cognitieve Deelhandelingen van het Rekenen (CDR). Handleiding & testprotocol [Cognitive Developmental skills in Arithmetics. Manual & testprotocol].* Herenthals: VVL.

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement, 6,* 440-458. https://doi.org/10.1177/0146621603258786

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

Dworak, E., Revelle, W., Doebler, P., & Condon, D. (2021). Using the International Cognitive Ability Resource as an open source tool to explore individual differences in cognitive ability. *Personality and Individual Differences*, *169*, Article 109906. https://doi.org/10.1016/j.paid.2020.109906

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN review system for evaluating test quality.* Boom Uitgevers Amsterdam. Retrieved from https://psynip.nl/wp-content/uploads/2022/05/COTAN-review-system-for-evaluating-test-quality.pdf

Evers, A., Muniz, J., Bartram, D., Boben, D., Egeland, J., Fernandez-Hermida, J. R., … Urbanek, T. (2012). Testing practices in the 21st century Developments and European psychologists' opinions. *European Psychologist, 17,* 300-319. https://doi.org/10.1027/1016-9040/a000102

Fergusson, D. M., & Horwood, L. J. (1997). Gender differences in educational achievement in a New Zealand birth cohort. *New Zealand Journal of Educational Studies, 32,* 83–96.

Flanagan, D., & Alfonso, V. (2017). *Essentials of WISC-V Assessment*. John Wiley & Sons.

Flanagan, D. P., & McDonough, E. M. (2018). *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (Fourth Edition). Guilford Press.

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (pp. 1–7). Tilburg, The Netherlands: Tilburg University Press.

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

Gottfredson, L., & Saklofske, D. H. (2009). Intelligence: Foundations and issues in assessment. *Canadian Psychology/Psychologie canadienne, 50,* 183. https://doi.org/10.1037/a0016641

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

IBM Corp. (2022). *IBM SPSS Statistics for Windows* (Version 29.0). Armonk, NY: IBM Corp

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

Kanazawa, S. (2013). Why is intelligence associated with stability of happiness? *British Journal of Psychology, 105,* 316-337. https://doi.org/10.1111/bjop.12039

Kelley, K., & Lai, K. (2012). MBESS: MBESS. R package version 3.3.2. Retrieved from http://CRAN. R-project.org/package=MBESS

Koenig, K. A., Frey, M. C., & Detterman, D. K. (2008). ACT and general cognitive ability. *Intelligence, 36,* 153–160. https://doi.org/10.1016/j.intell.2007.03.005

Kort, W., Schittekatte, M., Compaan, E. L., Bosmans, M., Bleichrodt, N., Vermeir, G., Resing, W. & Verhaeghe, P. (2002). *Handleiding WISC-III*^{NL}. London, Amsterdam: The Psychological Corporation, NIP Dienstencentrum.

Kristjánsdóttir, D., & Zaiter, A. (2023). *Public domain intelligence tests: Psychometric properties of the Cog15 and ICAR16 cognitive ability scales* [Master thesis, Lund University]. Retrieved from https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9125503&fileOId=9125512

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of statistical software, 82,* 1-26. https://doi.org/10.18637/jss.v082.i13

Kwaliteitscentrum voor Diagnostiek (2023). *Beoordeling van de WISC-V-NL voor gebruik in Vlaanderen.* https://portaal.kwaliteitscentrumdiagnostiek.be/actueel/beoordeling-van-de-wisc-v-nl-voor-gebruik-in-vlaanderen/

Lenroot, R. K., & Giedd, J. N. (2006). Brain development in children and adolescents: Insights from anatomical magnetic resonance imaging. *Neuroscience and Biobehavioral Reviews, 30,* 718-729. https://doi.org/10.1016/j.neubiorev.2006.06.001

Loe, B., Sun, L., Simonfy, F., & Doebler, P. (2018). Evaluating an Automated Number Series Item Generator Using Linear Logistic Test Models. *Journal of Intelligence, 6*, 20.

Low, L. K., & Cheng, H. J. (2006). Axon pruning: An essential step underlying the developmental plasticity of neuronal connections. *Philosophical Transactions of the Royal Society B-Biological Sciences, 361,* 1531-1544. https://doi.org/10.1098/rstb.2006.1883

Magez, W. (2019). *CoVaT-CHC Basisversie: Cognitieve VaardigheidsTest volgens CHC-model.* Centrum voor Psychodiagnostiek Thomas More. https://drive.google.com/file/d/1u4sXWgn5J0hhU_lk8xkWpAX9vjDZe60c/view

Magez, W., Tierens, M., Van Huynegem, J., Van Parijs, K., Decaluwé, V., & Bos, A. (2015). *CoVaT-CHC Basisversie.* Antwerpen: Psychodiagnostisch Centrum (PDC), Thomas More Hogeschool

Marcoulides, K. M., & Raykov, T. (2019). Evaluation of variance inflation factors in regression models using latent variable model ing methods. *Educational and Psychological Measurement, 79,* 874-882. https://doi.org/10.1177/0013164418817803

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

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*, 388–392. https://doi.org/10.1177/0734282920958220

Merkle, E. C., Fan, J. & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika, 79,* 569–584. https://doi.org/10.1007/s11336-013-9376-7

Ministry of Education and Training. (2012). *Omzendbrief SO/2012/01*. (SO/2012/01). Retrieved from http://data-onderwijs.vlaanderen.be/edulex/document.aspx?docid=14370

Ministry of Education and Training. (2021). *Omzendbrief SO/2021/01.* (SO/2021/01). Retrieved from https://data-onderwijs.vlaanderen.be/edulex/document.aspx?docid=15865

Ministry of Education and Training. (2022). Voltijds gewoon secundair onderwijs: algemeen overzicht (xlsx, 10 bladen) (150 kB) [Microsoft Excel spreadsheet]. Microsoft Corporation.

Ministry of Education and Training. (s.d.). *Naar het voltijds gewoon secundair onderwijs.* Retrieved from https://onderwijs.vlaanderen.be/nl/ouders/naar-school/naar-het-secundair-onderwijs/naar-het-voltijds-gewoon-secundair-onderwijs

Na, S., & Burns, T. (2016). Wechsler Intelligence Scale for Children-V: Test review. *Applied Neuropsychology: Child*, *5*(2), 156–160. https://doi.org/10.1080/21622965.2015.1015337

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in ecology and evolution, 4,* 133-142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Nelissen, W., Schittekatte, M., Fontaine, J., & Rigolle, F. (In preparation). *Validation of the International Cognitive Ability Resource (ICAR) in a Flemish sample*.

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist*, *67*, 130–159. https://doi.org/10.1037/a0026699

Oakland, T., Douglas, S., & Kane, H. (2016). Top ten standardized tests used internationally with children and youth by school psychologists in 64 countries: A 24-year follow-up study. *Journal of Psychoeducational Assessment*, *34*(2), 166–176. https://doi.org/10.1177/0734282915595303

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41,* 673-690. https://doi.org/10.1007/s11135-006-9018-6

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

R Core Team. (2024). *R: A language and environment for statistical computing* (version 4.3.3) [Software]. Vienna, Austria. Retrieved from http://www.R-project.org/

Raykov, T. (1998). Coefficient alpha and composite reliability with interrelated nonhomogeneous items. *Applied Psychological Measurement, 22,* 375–385. https://doi.org/10.1177/014662169802200407

Revelle, W., Condon, D., Wilt, J., French, J., Brown, A., & Elleman, L. (2017). Web- and phone-based data collection using planned missing designs. In N. Fielding, R. Lee, & G. Blank (Eds.), *The SAGE Handbook of Online Research Methods* (2nd ed., pp. 578–595). Sage Publishing.

Revelle, W., Dworak, E., & Condon, D. (2020). Cognitive ability in everyday life: The utility of open-source measures. *Current Directions in Psychological Science*, *29*(4), 358–363. https://doi.org/10.1177/0963721420922178

Rosseel, Y. (2012). Lavaan: An R package for Structural Equation Modeling. *Journal of Statistical Software*, *48*, 1-36. https://doi.org/10.18637/jss.v048.i02

Roth, B., Becker, N., Romeyke, S., Schäfer, S., 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

Sackett, P. R., Lievens, F., Van Iddekinge, C. H., & Kuncel, N. R. (2017). Individual differences and their measurement: A review of 100 years of research. *Journal of Applied Psychology, 102,* 254–273. https://doi.org/10.1037/apl0000151

Schmidt, F. L. (2016). *The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 100 years of research findings.* ResearchGate. Available at https://doi.org/10.13140/RG.2.2.18843.26400

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

Seghers, M., Boone, S., & Van Avermaet, P. (2019). Classed patterns in the course and outcome of parent-teacher interactions regarding educational decision-making. *Educational Review, 73,* 417-435. https://doi.org/10.1080/00131911.2019.1662771

Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., … Giedd, J. (2006). Intellectual ability and cortical development in children and adolescents. *Nature, 440,* 676-679. https://doi.org/10.1038/nature04513

Shaw, M., Rights, J. D., Sterba, S. S., & Flake, J. K. (2023). r2mlm: An R package calculating R-squared measures for multilevel models. *Behavior Research Methods, 55,* 1942-1964. https://doi.org/10.3758/s13428-022-01841-4

Stinissen, J., Smolders, M., & Coppens-Declerck, L. (1975). Handleiding bij de Collectieve Verbale Intelligentietest voor derde en vierde leerjaar (CIT 3-4). Brussel: C.S.B.O.

Tikhomirova, T., Malykh, A., & Malykh, S. (2020). Predicting academic achievement with cognitive abilities: Cross-sectional study across school education. *Behavioral Sciences, 10:*158. https://doi.org/10.3390/bs10100158

The International Cognitive Ability Resource Team. (2014). *International Cognitive Ability Resource*. https://icar-project.com

Vilia, P. N., Candeias, A. A., Neto, A. S., Franco, M. S., & Melo, M. (2017). Academic achievement in physics-chemistry: the predictive effect of attitudes and reasoning abilities. *Front. Psychol. 8:*1064. https://doi.org/10.3389/fpsyg.2017.01064

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence-Second Edition (WASI-II)* [Database record]. APA PsycTests. https://doi.org/10.1037/t15171-000

Weiss, L., Saklofske, D., Holdnack, J., & Prifitera, A. (Eds.). (2019). *WISC-V: Clinical Use and Interpretation*. Elsevier.

Wrulich, M., Brunner, M., Stadler, G., Schalke, D., Keller, U., & Martin, R. (2014). Forty years on: Childhood intelligence predicts health in middle adulthood. *Health Psychology, 33,* 292–296. https://doi.org/10.1037/a0030727

Young, S., & Keith, T. (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

Young, S. R., Keith, T. Z., & Bond, M. A. (2019). Age and sex invariance of the International Cognitive Ability Resource (ICAR). *Intelligence, 77,* 101399. https://doi.org/10.1016/j.intell.2019.101399

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

Zorowitz, S., Chierchia, G., Blakemore, S. J., & Daw, N. D. (2023). An item response theory analysis of the matrix reasoning item bank (MaRs-IB). *Behavior Research Methods.* https://doi.org/10.3758/s13428-023-02067-8

Appendix A

Table A1

Specimen of Each Item Type of the Ch-ICAR (Loe et al., 2018; The International Cognitive Ability Resource Team, 2014)

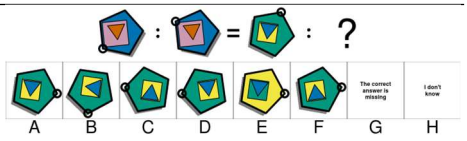
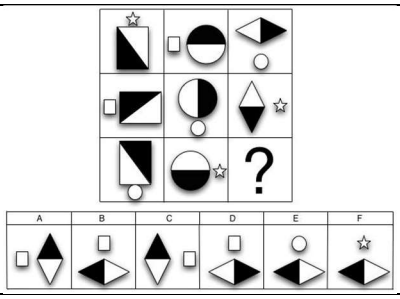
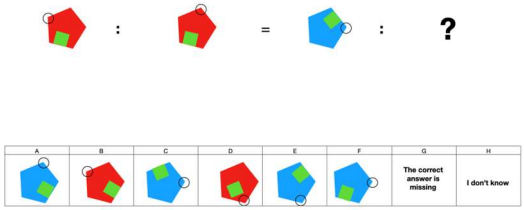
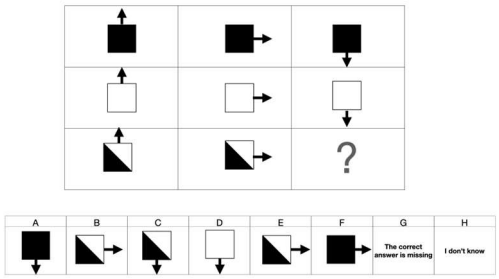
Verbal Reasoning	If the day after tomorrow is two days before Thursday then what day is it today? (1) Friday (2) Monday (3) Wednesday (4) Saturday (5) Tuesday (6) Sunday (7) None of these (8) I don't know	Figural Analogies	
Matrix Reasoning		Number Series	34,150,35,160,36,170, …

Table A2

Example Items of the Subtest Figural Analogies and Matrix Reasoning

Figural Analogies	Matrix Reasoning
	

Appendix B

Table B1

Overview Test Statistics Verbal Reasoning Items (Pilot Study, N = 46)

Cronbach's alpha full item set (16 items): .50 Cronbach's alpha final item selection (8 items): .65				
ICAR item number	**PC**	**Mean response time (in sec)**	**PC boys**	**PC girls**
VR18	.91	24	.91	.92
VR31	.56	42	.52	.58
VR39	.77	23	.90	.65
VR11	.12	19	.10	.13
VR14	.42	37	.50	.35
VR17	.37	41	.25	.48
VR42	.42	18	.40	.43
VR04	.10	22	.16	.04
VR19	.52	35	.53	.52
VR32	.43	39	.53	.35
VR09	.03	22	.00	.04
VR16	.21	37	.35	.09
VR36	.03	18	.06	.00
VR23	.26	19	.41	.14
VR26	.00	17	.00	.00
VR13	.15	40	.18	.14

Note. The numbers of the items refer to the numeration in the ICAR database (The International Cognitive Ability Resource Team, 2014). PC = proportion of correct responses in the full sample; PC boys = proportion of correct responses in boys subsample; PC girls = proportion of correct responses in girls subsample.

Table B2

Overview Test Statistics Matrix Reasoning Items (Pilot Study, N = 66)

Cronbach's alpha full item set (11 items): .63 Cronbach's alpha final item selection (8 items): .69				
ICAR item number	**PC**	**Mean response time (in sec)**	**PC boys**	**PC girls**
fig12043	.52	62	.64	.41
fig12044	.52	41	.56	.49
fig12047	.36	36	.41	.33
fig12046	.30	33	.33	.28
fig12053	.26	39	.30	.23
fig12045	.29	28	.30	.28
fig12048	.23	34	.33	.15

fig12054	.17	31	.22	.13
fig12056	.31	27	.33	.30
fig12055	.16	29	.22	.11
fig12050	.13	26	.11	.14

Note. The numbers of the items refer to the numeration in the ICAR database (The International Cognitive Ability Resource Team, 2014). PC = proportion of correct responses in the full sample; PC boys = proportion of correct responses in boys subsample; PC girls = proportion of correct responses in girls subsample.

Table B3

Overview Test Statistics Number Series Items (Pilot Study, N = 71)

Cronbach's alpha full item set (23 items): .88
Cronbach's alpha final item selection (8 items): .85

ICAR item number	PC	Mean response time (in sec)	PC boys	PC girls
8	.83	33	.83	.83
6	.90	20	.90	.90
2	.66	34	.79	.56
7	.89	18	.86	.90
1	.63	28	.79	.51
3	.51	35	.66	.39
4	.63	28	.79	.51
5	.52	28	.61	.44
10	.44	33	.52	.39
12	.49	28	.62	.40
9	.20	46	.19	.20
25	.22	45	.19	.24
19	.25	41	.23	.26
26	.32	41	.31	.33
27	.37	45	.46	.30
11	.09	56	.13	.06
33	.00	44	.00	.00
42	.09	49	.20	.03
13	.08	37	.05	.09
16	.10	39	.05	.12
44	.08	34	.11	.06
40	.06	38	.11	.03
28	.04	22	.06	.03

Note. The numbers of the items refer to the numeration in the ICAR database (The International Cognitive Ability Resource Team, 2014). PC = proportion of correct responses in the full sample; PC boys = proportion of correct responses in boys subsample; PC girls = proportion of correct responses in girls subsample.

Table B4

Overview Test Statistics Figural Analogies Items (Pilot Study, N = 56)

Cronbach's alpha full item set (20 items): .77
Cronbach's alpha final item selection (7 items): .83

ICAR item number	PC	Mean response time (in sec)	PC boys	PC girls
q18005	.36	68	.30	.41
q18019	.38	35	.31	.45
q18020	.44	28	.36	.52
q18006	.51	25	.52	.50
q18002	.26	24	.16	.36
q18001	.25	20	.16	.32
q18004	.25	36	.24	.25
q18009	.26	24	.13	.37
q18010	.31	19	.17	.42
q18003	.27	22	.13	.38
q18012	.29	28	.39	.19
q18011	.19	31	.26	.12
q18014	.15	20	.14	.16
q18018	.13	21	.14	.12
q18013	.33	18	.29	.38
q18017	.07	21	.05	.08
q18016	.24	19	.19	.29
q18015	.22	21	.19	.25
q18008	.05	16	.05	.04
q18007	.09	20	.15	.04

Note. The numbers of the items refer to the numeration in the ICAR database (The International Cognitive Ability Resource Team, 2014). PC = proportion of correct responses in the full sample; PC boys = proportion of correct responses in boys subsample; PC girls = proportion of correct responses in girls subsample.

Appendix C

Table C1

Included ICAR Items in the Children's ICAR

Subtest	ICAR item number
Verbal Reasoning	VR39, VR31, VR19, VR14, VR32, VR16, VR11, VR04
Matrix Reasoning	12044, 12043, 12056, 12046, 12053, 12048, 12055, 12050
Number Series	8, 5, 3, 12, 10, 27, 25, 13
Figural Analogies	q18005, q18020, q18019, q18006, q18015, q18009, q18016

Note. The numbers of the items refer to the numeration in the ICAR database (The International Cognitive Ability Resource Team, 2014)

Appendix D

Table D1

Item Fit Final Hierarchical IRT Model (Study 1, N = 820)

Item	S_ χ^2 Statistic	df	RMSEA	FDR adjusted p-value
MR1	12.5	15	.000	.719
MR2	24.4	13	.033	.105
MR3	20.2	14	.023	.258
MR4	12.9	14	.000	.632
MR6	11.8	15	.000	.719
MR7	8.4	15	.000	.906
VR1	15.3	13	.015	.433
VR2	23.9	11	.038	.060
VR3	16.7	13	.019	.382
VR4	28.1	12	.041	.048
VR5	35.0	14	.043	.020
VR6	17.0	14	.016	.431
VR7	16.5	14	.015	.433
VR8	17.2	13	.020	.363
NS1	19.2	12	.027	.215
NS2	26.8	13	.036	.060
NS3	10.9	10	.011	.509
NS4	35.9	10	.056	.002
NS5	16.6	9	.032	.169
NS7	14.0	13	.009	.509
NS8	12.5	13	.000	.601
FA1	11.2	11	.005	.546
FA2	17.7	11	.027	.215
FA3	19.3	11	.030	.169
FA4	18.7	12	.026	.215
FA5	11.1	14	.000	.719
FA6	30.0	14	.037	.052

Appendix E

Table E1

Correlations Between Academic Performance and SES Indicators (Study 1, N = 820)

		1.	2.	3.	4.	5.	6.	7.
1.	GPA	1						
2.	Mathematics	.86**	1					
3.	CDR	.37**	.40**	1				
4.	Home language	-.11**	-.10**	-.14**	1			
5.	Bursary	-.27**	-.27**	-.23**	.20**	1		
6.	Education parent 1	-.02	-.02	-.15**	.29**	.24**	1	
7.	Education parent 2	-.11**	-.10*	-.18**	.29**	.28**	.46**	1

Note. *<.05; **<.01.

Appendix F

Table F1

Internal Consistency of Ch-ICAR (sub)tests in CoVaT Sample (Study 2, N = 96)

(sub)test	McDonald's omega (ω)	Cronbach's alpha (α)
Verbal Reasoning	.60	.58
Matrix Reasoning	.52	.49
Number Series	.76	.74
Figural Analogies	.77	.73
Full Ch-ICAR	.84	.83

Table F2

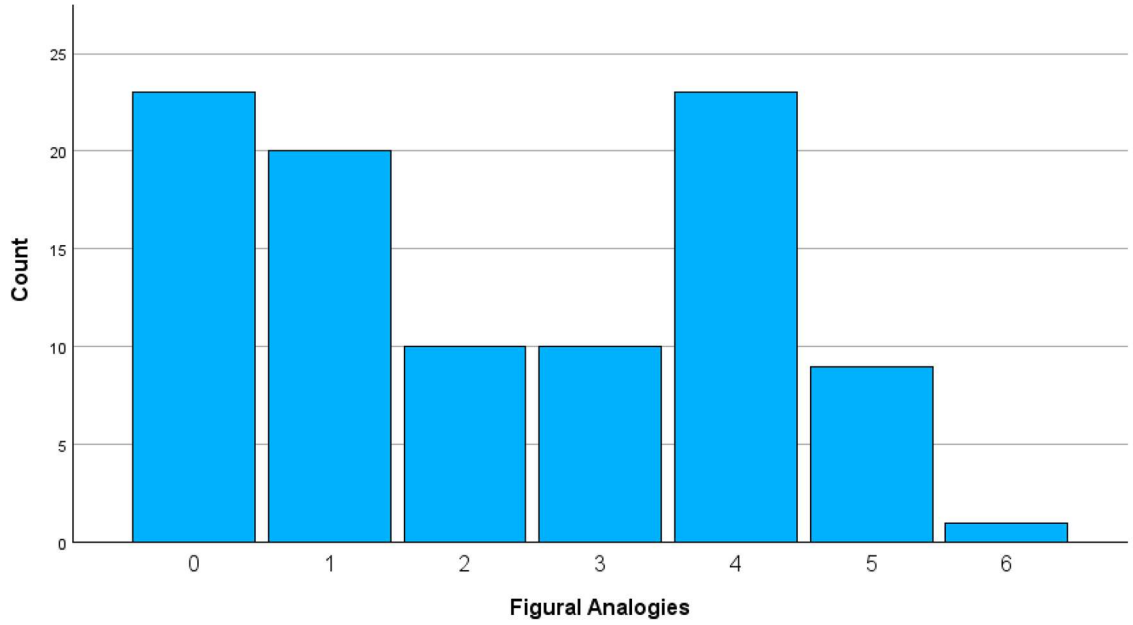
Internal Consistency of Ch-ICAR (sub)tests in RPM Sample (Study 2, N = 91)

(sub)test	McDonald's omega (ω)	Cronbach's alpha (α)
Verbal Reasoning	.66	.66
Matrix Reasoning	.31	.31
Number Series	.77	.75
Figural Analogies	.75	.70
Full Ch-ICAR	.81	.81

Appendix G

Figure G1

Distribution of Figural Analogies Scores (Study 2, CoVaT sample, N = 96)



Appendix H

Table H1

Proportion of Correct Responses per Item (Study 1 & Study 2)

Item	PC *(SD)* Study 1, *N* = 820	PC (*SD*) Study 2 CoVaT sample, N = 96	PC (*SD*) Study 2 RPM sample, N = 91
VR1	.83 *(.38)*	.89 *(.32)*	.78 *(.42)*
VR2	.73 *(.44)*	.84 *(.37)*	.87 *(.34)*
VR3	.48 *(.50)*	.53 *(.50)*	.52 *(.50)*
VR4	.38 *(.49)*	.44 *(.50)*	.41 *(.49)*
VR5	.43 *(.50)*	.43 *(.50)*	.47 *(.50)*
VR6	.20 *(.40)*	.21 *(.41)*	.35 *(.48)*
VR7	.20 *(.40)*	.29 *(.46)*	.33 *(.47)*
VR8	.24 *(.43)*	.30 *(.46)*	.25 *(.44)*
MR1	.38 *(.49)*	.44 *(.50)*	.47 *(.50)*
MR2	.49 *(.50)*	.73 *(.45)*	.71 *(.45)*
MR3	.24 *(.43)*	.31 *(.47)*	.32 *(.47)*
MR4	.30 *(.46)*	.35 *(.48)*	.33 *(.47)*
MR5	.30 *(.46)*	.49 *(.50)*	.32 *(.47)*
MR6	.29 *(.45)*	.36 *(.48)*	.46 *(.50)*
MR7	.17 *(.37)*	.35 *(.48)*	.21 *(.41)*
MR8	.14 *(.34)*	.18 *(.38)*	.18 *(.38)*
NS1	.82 *(.38)*	.87 *(.33)*	.89 *(.31)*
NS2	.51 *(.50)*	.56 *(.50)*	.55 *(.50)*
NS3	.56 *(.50)*	.67 *(.47)*	.70 *(.46)*
NS4	.42 *(.49)*	.44 *(.50)*	.55 *(.50)*
NS5	.46 *(.50)*	.59 *(.49)*	.68 *(.47)*
NS6	.19 *(.39)*	.20 *(.40)*	.20 *(.40)*
NS7	.25 *(.43)*	.21 *(.41)*	.42 *(.50)*
NS8	.17 *(.38)*	.24 *(.43)*	.33 *(.47)*
FA1	.45 *(.50)*	.42 *(.50)*	.59 *(.49)*
FA2	.49 *(.50)*	.51 *(.50)*	.67 *(.47)*
FA3	.48 *(.50)*	.49 *(.50)*	.68 *(.47)*
FA4	.52 *(.50)*	.49 *(.50)*	.66 *(.48)*
FA5	.08 *(.27)*	.11 *(.32)*	.13 *(.34)*
FA6	.20 *(.40)*	.20 *(.40)*	.24 *(.43)*
FA7	.15 *(.36)*	.21 *(.41)*	.26 (.44)