



Generational intelligence tests score changes in Spain: Are we asking the right question?

Roberto Colom^{*}, Luis F. García, Pei Chun Shih, Francisco J. Abad

Universidad Autónoma de Madrid, Spain

ARTICLE INFO

Keywords:

Generational intelligence changes

Cognitive test scores

Cohorts

Cross-sectional longitudinal design

ABSTRACT

Generational intelligence test score gains have been documented worldwide in the twentieth century. However, recent evidence suggests these increased scores are coming to an end in some world regions. Here we compare two cohorts of university freshmen. The first cohort ($n = 311$) was assessed in 1991, whereas the second cohort ($n = 349$) was assessed thirty years later (2022). These cohorts completed the same intelligence battery including eight standardized speeded and power tests tapping reasoning (abstract and quantitative), language (vocabulary, verbal comprehension, and verbal meanings), rote calculation, and visuospatial relations. The results revealed a global gain of 3.5 IQ points but also upward and downward changes at the test level. The 2022 cohort outperformed the 1991 cohort on reasoning (abstract and quantitative), verbal comprehension, and vocabulary, whereas the 1991 cohort outscored the 2022 cohort on rote calculation, visuospatial relations (mental rotation and identical figures), and verbal meanings. These findings are thought to support one key claim made by James Flynn: generational changes on the specific cognitive abilities and skills tapped by standardized tests should be expected without appreciable or substantive changes in the structure of the intelligence construct identified within generations. This main conclusion is discussed with respect to theoretical causal implications putatively derived from current intelligence psychometric models.

1. Introduction

Meta-analyses have demonstrated sustained generational intelligence tests score changes across the twentieth century worldwide (Trahan, Stuebing, Fletcher, & Hiscock, 2014; Wongupparaj, Kumari, & Morris, 2015). The meta-analysis by Pietschnig and Voracek (2015), based on data registered between 1909 and 2013 from 31 countries and 4 million people, revealed stronger gains for fluid (4.1 IQ points per decade) than for crystallized (2.1 IQ points per decade) ability. However, there is evidence that these increased scores are coming to an end in some world regions (Bratsberg & Rogeberg, 2018; Dutton, van der Linden, & Lynn, 2016; Flynn & Shayer, 2018; Hegelund et al., 2021).

The scientific community is still discussing the causes of these generational intelligence score changes. Education, nutrition/health care, thinking habits, cultural changes, family size, increased familiarity with standardized testing situations, or heterosis are among the nominated potential causes (Dickens & Flynn, 2001; Eppig, Fincher, & Thornhill, 2010; Lynn, 2009, 2013; Mingroni, 2007; Neisser, 1997; Schooler, 1998; Teasdale & Owen, 1987; Woodley, Te Nijenhuis, Must,

& Must, 2014; Zajonc & Mullally, 1997). All or some of these causes might (a) be now differentially active in different world regions, and (b) belong to the past in some of these regions. Nevertheless, the fact is that we are far from having unarguable explanations of the documented generational changes (Flynn, 2018).

One important topic relates to the question of whether or not the observed score changes involve real upward and downward trends at the construct level, namely, if the messengers (the tests) are telling a genuine story regarding intelligence gains and losses around the globe and across time. Some researchers argue that only changes at the construct level may be accepted as real (Gignac, 2015; te Nijenhuis & Van Der Flier, 2013), whereas others sustain that what happens within generations might be invalid for comparing generations (Dickens & Flynn, 2001). As underscored by Flynn (2007, p.10) “asking whether IQ gains are intelligence gains is the wrong question because it implies all or nothing cognitive progress (...) To assess cognitive trends, we must dissect intelligence into solving mathematical problems, interpreting the great works of literature, finding on-the-spot solutions, assimilating the scientific worldview, critical acumen, and wisdom.” Moreover,

^{*} Corresponding author at: Facultad de Psicología, Universidad Autónoma de Madrid, Spain.

E-mail address: roberto.colom@uam.es (R. Colom).

“intelligence can act like a highly correlated set of abilities on one level (individual differences) and like a set of functionally independent abilities on another level (cognitive trends over time)” (Flynn, 2012, p. 28).

Here we combine this latter approach (cognitive abilities are correlated within generations, whereas these abilities may show functional autonomy between generations) with one putative implication of intelligence psychometric models regarding the explanation of the universally replicated positive manifold phenomenon (Warne & Burningham, 2019). Thus, for instance, hierarchical models of intelligence involve causal connections going downwards, from the highest to the lowest levels (Fig. 1).

There are several psychometric models that may provide accounts for the universal positive correlation among cognitive abilities (common cause, interconnected, and sampling models) and, to our knowledge, there have been two complementary approaches to test their likelihood beyond the statistical realm. The first approach explores the positive impact of treatments and training programs over specific cognitive abilities and its resonance on the higher-order (*g*) factor of intelligence (Protzko, 2017), whereas the second analyzed the negative impact of focal and chronic cortical lesions over specific cognitive abilities and its effect on *g* (Protzko & Colom, 2021). So far, both approaches led to conclude that the observed positive and negative outcomes do show local instead global effects.

This combined perspective is the main framework for the research reported here addressing generational intelligence changes in Spain (Fig. 1). Specifically, we expect distinguishable changes between generations on intelligence tests tapping several cognitive abilities even when they are correlated within generations. For testing this expectation, we recover the raw scores on eight intelligence tests completed by university freshmen in 1991 and another group who completed the same battery thirty years later (2022). Therefore, we use here a cross-sectional longitudinal approach (Gravetter & Forzano, 2018). Firstly, we will show how test scores behave within generations. Secondly, we will compute generational changes from these scores. Finally, we will discuss the implications of the observed findings for obtaining answers to these main questions: do tests scores change in an orchestrated way across generations? Do they show functional autonomy,¹ as argued by Flynn (2007, 2012)?

2. Method

2.1. Participants

Two cohorts are considered here. The first cohort comprises 311 university freshmen assessed in 1991, whereas the second cohort includes 349 students assessed 30 years later (2022). All were tested within the first weeks of their first course administering the same intelligence test battery. For the 1991 cohort there were almost no missing values (99.4% completed all subtests). 80.5% of the 2022 cohort completed all subtests and the average sample size across subtests was 324, ranging from 306 to 332 depending on the subtest. The study was approved by the Ethics Committee² and participants gave informed written consent.³ The mean age of the 2022 cohort was 18.3 years (SD = 1.1) and 83% were female. Age and sex data were unavailable for the 1991 cohort.

¹ Flynn (2007, pp. 18 and 65) explains what functional autonomy means: “at any particular time, factor analysis will extract *g* and intelligence appears unitary. Over time, real-world cognitive skills assert their functional autonomy and swim freely of *g* and intelligence appears multiple (...) social trends show that various cognitive skills are largely functionally independent of one another”.

² CEI-122-2484.

³ There was a very small number of students (3%) that chose not to complete the battery.

We recovered evidence supporting the presumption that both cohorts might be meaningfully compared.⁴ First, the proportion of the population attending universities in Spain because candidates passed the state examination has increased since the 90s, moving from 33% to 55.1%. Second, the pass rate for the exam has increased from 72.8% to 90.2%. Third, the cut score for admission to the Faculty of Psychology has increased accordingly from 5.3 to 7.9. These facts seem consistent with grade inflation (Roth et al., 2015). Obtaining better grades seems easier now than one generation ago (please see limitations section below for further arguments regarding the comparability of both cohorts).

2.2. Intelligence battery

The intelligence battery was composed of eight tests chosen to tap reasoning (abstract and quantitative), verbal ability (vocabulary, verbal comprehension, and verbal meanings), spatial relations (rotation of solid figures and identical figures), and rote calculation. The main features of the administered test are described next (the Appendix depicts examples).

2.2.1. Reasoning

2.2.1.1. DAT-AR (abstract reasoning). This is a series power test based on abstract figures. This test comprises fifty items. Each item includes four figures following a given rule, and the participant must choose one of five possible alternatives properly completing the series. The score is the total number of correct responses. Completion time = 25 min (TEA, S.A, 1979a).

2.2.1.2. Monedas-2 (quantitative reasoning). This is a quantitative reasoning power test comprising 40 items. The items are based on the combination of the size of a series of coins (large, medium, and small), the digits put inside the coins to specify the number of them that the participants must consider, and some numerical operations to make the necessary calculations to arrive at a given response (adding, subtracting, and so forth). Only one alternative is correct. The score obtained is the total number of correct responses. Completion time = 12 min (Seisdedos, 1978).

2.2.2. Verbal ability

2.2.2.1. PMA-V (vocabulary). This is a synonym speeded test that comprises 50 items. The meaning of four alternative letters must be evaluated against a given letter that serves as model. Only one alternative is correct. The score is the total number of correct responses. Completion time = 4 min (TEA, S.A, 1979b).

2.2.2.2. Verbal comprehension. This power test includes thirty items. Each item is based on the proper understanding of a given sentence that must be compared with three alternatives of which only one carries the same thought. The score obtained is the total number of correct responses. Completion time = 15 min (Manziona, 1978).

2.2.2.3. Verbal meanings. This speeded test includes 20 items. Each item is composed of a sentence with an uppercase word. Participants must choose among five words the one that keeps the meaning of the sentence. Items increase their complexity across the test. The score obtained is the total number of correct responses. Completion time = 4 min (Manziona, 1978).

⁴ INE (Instituto Nacional de Estadística); Ministerio de Educación y Formación Profesional.

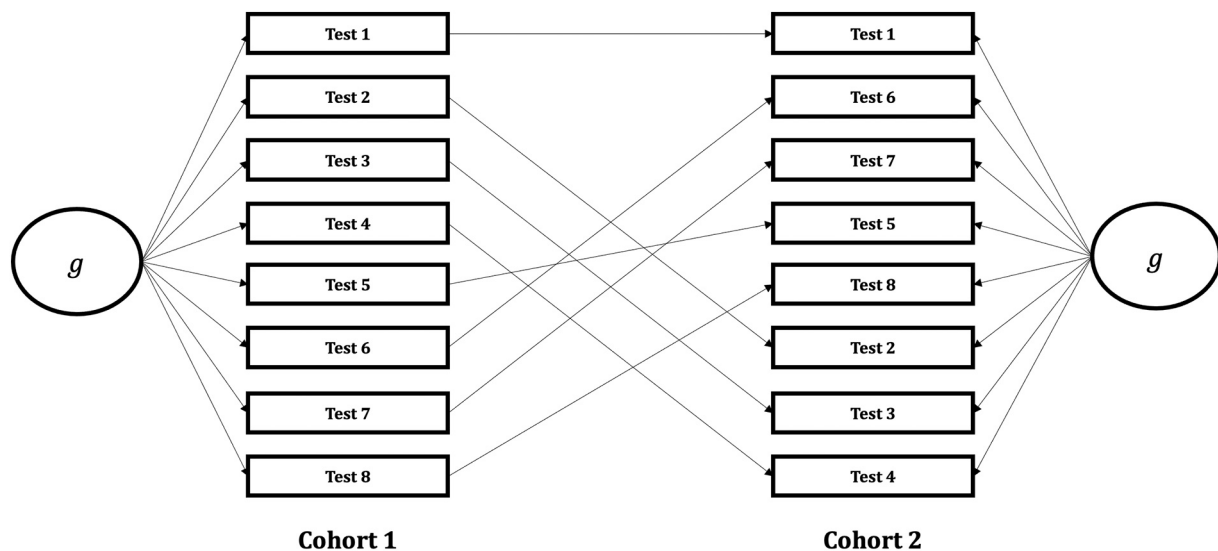


Fig. 1. General framework of the present study. Cognitive abilities tapped by several standardized tests can be similarly organized within generations (cohort 1 and cohort 2). However, across generations (cohort 1 versus cohort 2) putative upward and downward tests scores changes (arrows from cohort 1 to cohort 2) may or may not obey the organization rules active within generations. Common cause models of intelligence imply that changes at lower levels of the hierarchy (tests) will not be necessarily reflected at the higher levels across generations.

2.2.3. Spatial relations

2.2.3.1. Rotation of solid figures. This speeded test comprises 21 items. Each item includes a model figure, and five alternatives must be evaluated against it. The participant must evaluate which alternative can be rotated within a 3D space to fit the model figure. Only one alternative is correct. The score is the total number of correct responses. Completion time = 5 min (Yela, 1969).

2.2.3.2. Identical figures. This speeded test includes 25 items. Each item comprises a model figure and five alternatives of whom only one is identical to the model figure. The figures increase their complexity across the test. The score obtained is the total number of correct responses. Completion time = 3 min (Manziane, 1978).

2.2.4. Rote calculation

2.2.4.1. PMA-N (rote calculation). This is a calculation speeded test that comprises 70 items. The participant must simply evaluate if a given sum is correctly or badly solved. For instance, $16 + 38 + 45 = 99$? The score is the total number of correct Responses. Completion time = 6 min (TEA, S.A., 1979b).

The 2022 cohort also completed the RAPM (Raven Advanced Progressive Matrices) test (Set II). Each of the thirty-six items comprises a matrix figure with three rows and three columns. Among eight possible alternatives the one completing the 3×3 matrix figure must be chosen. Completion time = 40 min.

2.3. Procedure

The tests were administered in two sessions in groups of no more than forty students. In the first session, they completed the following tests: DAT-AR, PMA-V, and Monedas-2. In the second session, they completed the following tests: PMA-N, Verbal Meanings, Identical Figures, Rotation of Solid Figures, and Verbal Comprehension. The 2022 cohort completed an online version of the RAPM.

3. Analyses

Firstly, we analyze the psychometric behavior of the administered

tests within cohorts. For this purpose, the descriptive statistics were obtained. Afterwards, we computed a factor analyses (principal axis factoring) followed by an oblique Promax Rotation (Fabrigar et al., 1999). Next, using the output obtained from this latter computation we calculated a single factor score from the first unrotated principal factor summarizing the scores obtained by both cohorts. The number of factors was assessed by parallel analysis, based on principal components and using the mean of random eigenvalues as comparison criteria (Lim & Jahng, 2019). Random data were generated by sampling with replacement. We also used Exploratory Graph Analysis (Golino & Demetriou, 2017), which have been shown to be an accurate approach for estimating dimensionality in the presence of difficult conditions (i.e., high factor correlations, low factor loadings, few variables per factor). For the EGA, we apply the graphical least absolute shrinkage and selection operator (i.e., GLASSO) as a regularization technique to deal with spurious connections, and the Louvain clustering algorithm. bootEGA was used to assess the stability of the EGA dimensionality estimates and subtest factor assignments across many bootstrap samples.

Secondly, we compared both cohorts across the completed tests for obtaining a general picture of how their scores changed from 1991 to 2022. This approach also involved the computation of percentiles after the obtained raw scores in 1991 and 2022 to know whether there is some pattern regarding the region of the distribution of scores where cohort differences are more evident. For effect size measures (i.e., Cohen's d and correlations), bootstrap resampling was employed to estimate the confidence intervals, given the significant deviation of some variables' distribution from normality (within-cohort).

Finally, we computed MGCFA (Multi-Group Confirmatory Factor Analysis) for obtaining an answer to the question of whether generational changes at the lower level of the intelligence hierarchy (test level) involve changes of some sort at the higher level (g) (Wicherts et al., 2004). Three models (unidimensional, and two bifactor models) are considered as possible baseline models. In the first bifactor model, two group factors are modeled, whereas in the second bifactor model only one group factor for tests based on language (vocabulary, verbal comprehension, and verbal meanings) is considered. These models are fitted separately for each cohort and, for the chosen model, we sequentially test a series of nested models, progressively constrained (e.g., Benson, Beaujean, & Taub, 2015; Luong & Flake, 2022): (1) the baseline or configural invariance model, in which the same structure is

specified for each group but loadings and intercepts are let to vary across cohorts, (2) metric invariance or invariance of factor loadings, in which loadings are constrained to be equal across cohorts, and (3) scalar invariance or invariance of intercepts, in which intercepts are constrained to be equal across cohorts. Metric invariance implies that tests are equally related with the constructs across cohorts, which provides support for constructs having the same interpretation within-cohort. Total Scalar invariance is required for cohort comparisons in observed tests scores, which provides support for interpretation of between-cohort differences. If metric or scalar invariance are not tenable, then observed performance differences would be dependent on test specificity (subtest specific differences on loadings and intercepts across groups) and would not be totally due to the tapped constructs (factor means). For instance, for a case in which there are no group differences in latent factors mean and variances, if subtest loadings or intercepts differ, there will be a mean difference at the subtest level.

For model identification, factor variances (means) are always set to 1 (0) in the first cohort, and in the second cohort when all loadings (intercepts) are set free across cohorts. For parameter estimation, we use the robust maximum likelihood estimator (MLR) to obtain robust (Huber-White) standard errors and tests statistics, given the observed deviations from normality, and full-information maximum likelihood algorithm to deal with missing data. Model fit is assessed using the Yuan-Bentler chi-square statistic, the root mean square error of approximation (RMSEA), and the comparative fit index (CFI). CFI and RMSEA are based on the Yuan-Bentler chi-square. Values close to 0.95 for CFI and below 0.06 for RMSEA suggest a good fit (Hu & Bentler, 1999). To evaluate model fit relevant changes, we consider a decrease of CFI larger than 0.01 criteria, proposed by Cheung and Rensvold (2002), which is a commonly used criterion (Putnick & Bornstein, 2016). We also consider the Akaike Information Criterion (AIC), which has a relative interpretation, with smaller values indicating better fit. When results suggest lack of invariance, modification indices (MIs) are used sequentially to relax the equality constraints and to detect sources of invariance (which is known as a backward-selection approach; see Luong & Flake, 2022). The parameters for the subtest with the highest MI are set free across all the groups. This step is repeated until improvement in model fit is

satisfactory, according to the CFI decrease cutoff, or modification indices no longer indicating significant improvements in model fit (MI < 3.84, associated with a significance level of 0.05, in the one degree of freedom chi-square distribution). Note that this strategy can provide a partial measurement invariance model, in which the cohorts can be compared on latent (but not observed) means. On the other hand, for comparisons to be conceptually justified a majority of variables should be invariant (Luong & Flake, 2022).

Analyses were performed using SPSS 26 and R (R Core Team, 2021), the packages lavaan (Rosseel, 2012), semTools (Jorgensen, Pornprasertmanit, Schoemann, & Rosseel, 2022), EGAnet (Golino & Christensen, 2022) and BootES (Kirby & Gerlanc, 2013).

4. Results

The descriptive statistics are shown in Table 1. The means and SDs, range of scores, skewness and kurtosis values for both cohorts, along with the effect size representing the average difference (d) and the IQ equivalent ($d \times 15$) are shown in the bottom half of Table 1. The top half shows the correlation matrices for the 1991 cohort (above the diagonal) and for the 2022 cohort (below the diagonal). Reliability values obtained from the tests' manuals are shown at the diagonal. Note that correlation values are strikingly similar for both cohorts.

After examining the correlation matrices in Table 1, we assessed the suitability of the data for factor analysis. Both cohorts showed significant correlations between variables, as indicated by the Bartlett's test of sphericity ($p < 0.001$). The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was moderate for both cohorts (0.771 and 0.742, respectively). These results suggest that the datasets were suitable for factor analysis. Then, the dimensionality was analyzed. Parallel analysis suggested two factors for the 1991 cohort and one factor for the 2022 cohort, whereas EGA suggested two factors for both the cohorts (in 79.4% and 55.2% of the bootstrap samples, respectively). Afterwards, we computed, within cohorts, a principal axis factoring (PAF) followed by a Promax rotation relying on the recommendations considered in the Fabrigar et al. (1999) seminal article. The obtained outputs from the factor analysis are shown in Table 2. The second column of Table 2

Table 1

Correlation matrices from cohorts 1991 and 2022, descriptive statistics, effect sizes (d) and IQ points. Reliability values at the diagonal (a) Spearman-Brown, (b) Split-Half, (c) Test-retest (1 yr.). N.A. = not available.

1991 (Above) 2022 (Below)	Quantitative reasoning	Vocabulary	Abstract reasoning	Verbal comprehension	Mental rotation	Rote calculation	Verbal meanings	Identical figures
Quantitative Reasoning	0.94(a)	0.29**	0.51**	0.20**	0.36**	0.42**	0.25**	0.21**
Vocabulary	0.32**	0.91(b)	0.29**	0.28**	0.16**	0.32**	0.28**	0.08
Abstract Reasoning	0.48**	0.15**	0.91(b)	0.19**	0.30**	0.28**	0.26**	0.24**
Verbal Comprehension	0.31**	0.30**	0.24**	0.66(c)	0.05	0.09	0.35**	0.02
Mental Rotation	0.30**	0.19**	0.31**	0.23**	0.87(a)	0.22**	0.10	0.24**
Rote Calculation	0.37**	0.15**	0.17**	0.07	0.11	0.99(b)	0.17**	0.19**
Verbal Meanings	0.21**	0.30**	0.22**	0.21**	0.10	0.04	N.A.	0.11
Identical Figures	0.30**	0.17**	0.20**	0.09	0.20**	0.09	0.20**	0.90(c)
Mean 1991	20.80	28.01	36.88	15.30	9.79	19.90	14.50	14.31
SD 1991	6.49	6.78	6.88	3.74	4.85	7.56	2.19	2.57
Min.-Max.	2-34	8-49	6-49	0-27	1-21	1-49	4-19	2-21
Skewness	-0.37	0.04	-1.06	-0.03	0.26	0.77	-0.77	-1.03
Kurtosis	-0.31	0.10	1.74	0.40	-0.82	1.38	1.38	2.37
N	311	310	311	311	310	311	310	310
Mean 2022	21.98	29.60	40.50	16.85	8.52	15.53	13.86	13.66
SD 2022	6.67	5.95	4.37	3.09	3.59	4.92	2.15	2.26
Min.-Max.	5-40	14-45	26-50	9-26	3-21	5-34	7-18	6-21
Skewness	-0.04	0.03	-0.76	-0.05	0.88	0.53	-0.56	-0.41
Kurtosis	-0.28	-0.17	0.38	-0.07	0.57	0.62	0.24	1.34
N	332	331	330	325	306	320	325	323
Effect Size (d)	+0.18**	+0.25**	+0.63**	+0.45**	-0.30**	-0.69**	-0.29**	-0.27**
IQ points	+2.7	+3.7	+9.5	+6.8	-4.5	-10.3	-4.4	-4.0

Note. ** indicates that statistic (i.e., correlation, d) 95% confidence interval does not include the zero-value. Bias-corrected and accelerated (BCa) bootstrap intervals were computed (1000 bootstrap samples).

Table 2
Factor matrix (PAF-Promax Rotation).

Tests	First unrotated principal factor	Structure matrix	
	FUPF (1991/2022)	F1 (1991/ 2022)	F2 (1991/ 2022)
Quantitative Reasoning	0.73/0.76	0.75/0.79	0.45/0.42
Vocabulary	0.49/0.56	0.41/0.31	0.51/0.79
Abstract Reasoning	0.65/0.55	0.65/0.61	0.42/0.23
Verbal Comprehension	0.38/0.43	0.21/0.35	0.61/0.40
Mental Rotation	0.44/0.40	0.50/0.40	0.19/0.24
Rote Calculation	0.50/0.35	0.52/0.37	0.31/0.17
Verbal Meanings	0.44/0.35	0.31/0.25	0.56/0.39
Identical Figures	0.31/0.32	0.35/0.30	0.12/0.23
Variance explained by the common factor 26%/23.4%		r F1, F2 1991 = 0.54 r F1, F2 2022 = 0.46	

depicts the factor loadings on the first unrotated principal factor (FUPF) for both cohorts, whereas the third and fourth columns show the factor loadings on the obtained two rotated factors (F1 and F2). The rank order of the factor loadings on the FUPF of the completed intelligence tests is closely similar for both cohorts. Reasoning (quantitative and abstract), along with verbal comprehension and vocabulary show the highest loadings, whereas identical figures, verbal meanings, and rote calculation show the lowest loadings. Also, the amount of common variance explained by this FUPF is similar for the 1991 cohort (26%) and for the 2022 cohort (23.4%). Regarding the rotated factors, the structure matrix groups the completed tests similarly in both cohorts. The first factor summarizes reasoning (quantitative and abstract), spatial relations (mental rotation and identical figures), and rote calculation. The second factor summarizes the tests based on language (vocabulary, verbal comprehension, and verbal meanings). Finally, the correlation between these two factors is 0.54 for the 1991 cohort and 0.46 for the 2022 cohort. Noticeably, the same aggrupation was obtained for EGA, being quite stable across the bootstrap samples (when two clusters were extracted, the most common solution was replicated in the 99% of the cases for the 1991 cohort, and in 82% of the cases in the 2022 cohort).

In the next step we computed a factor analysis for obtaining weighted general scores at the individual level in both cohorts combined. These factor scores were transformed for clarity to the scale with a mean of 100 and a SD of 15. The 2022 values were: Mean = 101.8 and SD = 13.5. The 1991 values were: Mean = 98.3 and SD = 16.1. Therefore, the average difference was 3.5 IQ points favoring the 2022 cohort ($p = 0.006$). The analyses also revealed a statistically significant difference in the obtained standard deviation values: the 1991 cohort shows greater heterogeneity of scores (16.1) than the 2022 cohort (13.5).

Now we focus on the test level. As shown in Table 1, the 2022 cohort outperforms the 1991 cohort on abstract reasoning (+9.5 IQ points), verbal comprehension (+6.3 IQ points), vocabulary (+3.7 IQ points), and quantitative reasoning (+2.7 IQ points), whereas the 1991 cohort outscores the 2002 cohort on rote calculation (−10.3 IQ points), mental rotation (−4.5 IQ points), verbal meanings (−4.5 IQ points), and

identical figures (−4 IQ points). Therefore, the average advantage shown by the 2022 cohort was concentrated on the intelligence tests showing the highest factor loadings on the FUPF, while the average advantage shown by the 1991 cohort relied on the intelligence tests with the lowest factor loadings (Table 2). It is noteworthy that these average differences at the test level show remarkable heterogeneity, ranging from −10.3 IQ points to +9.5 IQ points, which might raise reservations with respect to the informative nature of the general estimation provided above and based on the obtained common factor.

Table 3 depicts the correspondence between raw and percentile scores (from P10 to P90), along with the difference between cohorts within the considered percentile bands. For most tests, these difference values do not reveal any interesting pattern. However, there is something that deserves attention with respect to abstract reasoning and rote calculation (rows highlighted in grey on Table 3). Remember that rote calculation showed the largest average advantage of the 1991 cohort (−10.3 IQ points), whereas abstract reasoning showed the largest average advantage of the 2022 cohort (+9.5 IQ points). Interestingly, the 1991 advantage on rote calculation is mainly concentrated in the highest and mean regions of the distribution of scores, while the 2022 advantage on abstract reasoning is mainly concentrated in the lowest and mean regions of the distribution.

Next, we computed the correlation between RAPM scores and the scores obtained on the eight tests completed by the 2022 cohort. These were the obtained r values: 0.47 (DAT-AR, abstract reasoning), 0.45 (Monedas-2, quantitative reasoning), 0.33 (Rotation of solid figures), 0.29 (Verbal comprehension), 0.25 (Identical figures), 0.24 (Verbal meanings), 0.17 (Rote calculation), and 0.16 (Vocabulary). Therefore, the highest correlations with the RAPM are concentrated on the reasoning tests, DAT-AR and Monedas-2, classified as power tests. The mean raw score on the RAPM was 23.04 and the standard deviation was 4.25.

Finally, regarding the invariance analyses Table 4 shows fit values for all the tested potential baseline models in each cohort. The unidimensional model did not fit the data, whereas bifactor models showed good fit ($CFI > 0.95$; $RMSEA < 0.06$). We retained the bifactor model with one group factor (BF1) since (a) in the 1991 cohort, BF1 and BF2

Table 4
Fit values for the baseline models (1991 and 2022 cohorts).

Model	X ²	df	CFI	RMSEA	AIC
<i>1991 cohort</i>					
Unidimensional	63.684	20	0.882	0.083	6760.98
BF1: Bifactor 1	23.190	17	0.984	0.033	6726.85
BF2: Bifactor 2	19.301	12	0.988	0.034	6730.37
<i>2022 cohort</i>					
Unidimensional	56.062	20	0.893	0.077	6045.82
BF1: Bifactor 1	32.461	17	0.953	0.056	6028.22
BF2: Bifactor 2	19.372	12	0.978	0.045	6024.59

Note. BF1: bifactor model with one group factor; BF2: bifactor model with two group factors.

Table 3
Percentile scores (P10 to P90) for the 1991 and the 2022 cohorts (1991/2022) on the eight tests and difference on the obtained raw scores within percentiles (diff.). Lines highlighted in bold type show the most extreme instances of progressively reduced differences (abstract reasoning) and increased differences (rote calculus) between cohort along the distribution of scores.

Tests	P10 1991/2022 (diff.)	P20	P30	P40	P50	P60	P70	P80	P90
Quantitative Reasoning	12/13 (+1)	15/16 (+1)	17/18 (+1)	19/20 (+1)	21/22 (+1)	24/24 (0)	25/26 (+1)	27/27 (0)	29/31 (+2)
Vocabulary	20/21 (+1)	23/24 (+1)	24/26 (+2)	26/29 (+3)	28/30 (+2)	30/31 (+1)	32/33 (+1)	34/34 (0)	37/37 (0)
Abstract Reasoning	27/35 (+8)	33/37 (+4)	34/39 (+5)	36/40 (+4)	38/41 (+3)	39/42 (+3)	41/43 (+2)	43/44 (+1)	45/46 (+1)
Verbal Comprehension	11/13 (+2)	12/14 (+2)	13/14 (+1)	14/16 (+2)	15/17 (+2)	16/18 (+2)	17/18 (+1)	18/19 (+1)	20/21 (+1)
Mental Rotation	4/4 (0)	5/5 (0)	6/6 (0)	8/7 (−1)	9/8 (−1)	11/9 (−2)	13/10 (−3)	15/11 (−4)	16/14 (−2)
Rote Calculation	12/10 (−2)	14/11 (−3)	15/13 (−2)	18/14 (−4)	19/15 (−4)	21/16 (−5)	22/18 (−4)	25/19 (−6)	30/22 (−8)
Verbal Meanings	12/11 (−1)	13/12 (−1)	14/13 (−1)	14/13 (−1)	15/14 (−1)	15/15 (0)	16/15 (−1)	16/16 (0)	17/16 (−1)
Identical Figures	11/11 (0)	13/12 (−1)	13/13 (0)	14/13 (−1)	15/14 (−1)	15/14 (−1)	16/15 (−1)	16/15 (−1)	17/16 (−1)

models showed similar fit indexes, but in the BF2 model only two out of five loadings on the reasoning factor were significant ($p < 0.05$) and (b) in the 2022 cohort, the BF2 solution was not admissible (i.e., some error variances were negative) and no loading on the reasoning factor was significant ($p > 0.05$).

Table 5 summarizes the fit indexes for all the invariance tested models (BF1 baseline model). The configural invariance model shows excellent fit, indicating that the hypothesized model adequately describes the data. When factor loadings are constrained to be equal across cohorts, fit significantly decreases ($\chi^2 = 37.733$; $df = 9$; $p < 0.001$), with a CFI decrease from 0.968 to 0.931. Interestingly, modification indices show that loadings on the general factor differed for the Rote Calculation and Abstract Reasoning tests. A partial metric invariance model including these modifications show a delta CFI decrease smaller than 0.01 regarding the original configural model (i.e., CFI decrease from 0.968 to 0.961) and lower AIC. Thus, this partial metric invariance model is considered as the referent for subsequent invariance analyses. When the intercepts for the remaining six tests are constrained, we find a significant decrease of fit ($\chi^2 = 139.95$; $df = 4$; $p < 0.001$), with a decrease of CFI from 0.961 to 0.845. Thus, scalar invariance is not tenable. Subsequent analyses show that only the intercepts for Vocabulary, Verbal Comprehension and Quantitative Reasoning are invariant, meaning that there is a large percentage of non-invariant tests (62.5%), and, therefore, variations in raw test scores between cohorts should not be interpreted as evidence of differences in latent abilities or traits (i.e., the general or the verbal factors). Instead, such differences may be partially attributed to more specific abilities required at the subtest level.

Table 6 shows parameters obtained for the partial scalar invariance bifactor model. From this output, it could be concluded that there has been a gain in the latent factors (for g , $d = 0.21$; for Verbal, $d = 0.68$). For the g factor, this gain is equivalent 3.2 IQ points. However, we must consider that, due to the large percentage of non-invariant tests noted above, the identification of the non-invariant set and the latent differences can be non-reliable. We only have three scalar invariant subtests (Quantitative Reasoning, Vocabulary and Verbal Comprehension), and, indeed, only one of them (Quantitative Reasoning) is a pure indicator of g .

Fig. 2 depicts two alternative views of the observed results, which are hardly distinguishable by statistical fit (please see Table 5). These figures show the standardized latent mean differences (Cohen's d) for the g and the Verbal factors, according to each model. They also show the difference of intercepts for each subtest. For the partial metric invariance model, all the intercepts (8) differ, and there is no difference due to the latent factors. On the contrary, for the partial scalar invariance model, there are some differences in latent factors, but it also includes five intercepts differing across cohorts. Both models differ only on one degree of freedom, so we consider that there is no serious reason for choosing the last. Indeed, we think that the most parsimonious model, given the pattern of increases and decreases in test performance, should be the partial metric invariance model.

Table 5

Fit indexes for all the invariance tested models.

Model	χ^2	df	CFI	RMSEA	AIC
Configural	55.817	34	0.968	0.046	12,755.08
Metric ¹	92.479	43	0.931	0.061	12,772.87
Partial metric ²	68.352	41	0.961	0.047	12,753.31
Scalar ³	160.263	45	0.845	0.089	12,831.93
Partial scalar ⁴	69.929	42	0.959	0.047	12,753.09

Note. ¹All loadings are constrained; ²All loadings are constrained, except those for Rote Calculation and Abstract Reasoning; ³All loadings and intercepts are constrained, except those for Rote Calculation and Abstract Reasoning; ⁴All the loadings are constrained, except those for Rote Calculation and Abstract Reasoning, but only Quantitative Reasoning, Vocabulary and Verbal Comprehension intercepts are constrained.

5. Discussion

5.1. Summary of the main findings

Here we have shown that, globally, there are still generational intelligence gains in Spain. The weighted factor scores for the 1991 and 2022 cohorts revealed an average increase equivalent to 3.5 IQ points. Moreover, this increase was accompanied by reduced variance in the 2022 cohort, which suggests that individuals tested nowadays are more alike than those tested one generation ago.

Secondly, at the test level results were distinguishable. The 2022 cohort obtained higher scores on abstract reasoning (DAT-AR), verbal comprehension, vocabulary (PMA-V), and quantitative reasoning (Monedas-2), whereas the 1991 cohort outscored the recent cohort on rote calculation (PMA-N), mental rotation, verbal meanings, and identical figures. Therefore, the finding at this test level paints an interesting picture of upward and downward movements in two generations separated by the change of millennia (Fig. 3).

Thirdly, scores across the bell curve tell one thought-provoking story with respect to the cognitive tests showing the greatest average differences between cohorts, namely, abstract reasoning (DAT-AR) and rote calculation (PMA-N). Generational gains on abstract reasoning are concentrated in the lowest and medium areas of the curve, whereas the opposite pattern is observed for rote calculation. The losses on this latter test are mainly manifested in the upper half of the distribution, whereas the gains on abstract reasoning are especially potent in the lower half of the distribution. Specifically, the gains on the DAT-AR are progressively reduced as we move upwards in the distribution of difference scores, from +8 (percentile 10) to +3 (percentile 50) to +1 (percentile 90), whereas the losses are progressively reduced on the PMA-N as we move downwards in the distribution of difference scores, from -8 (percentile 90), to -4 (percentile 50) to -2 (percentile 10). For the remaining mental tests, difference scores are mainly homogeneous across the considered percentile bands.

The pattern observed for abstract reasoning (DAT-AR) in our university freshmen tested one generation apart (from 1991 to 2022), resembles the pattern reported for Spain by Colom et al. (2005a) after the comparison of cohorts comprised by 7 years old children tested one generation apart (from 1970 to 1999). The children tested in 1999 showed an average advantage of +9.7 IQ points, but the gains were mainly concentrated in the lower and medium halves of the intelligence distribution (+9 P1, +6 P25, +4 P45, +3 P85, and +1 P99). Therefore, the change in difference scores across the bell curve for these children followed the same trend reported here for the DAT-AR in university freshmen from Spain. Remember that the average gain in this later case was +9.5 IQ points, or 3 IQ points per decade.

In summary, global gains are found in the registered data, but these should be evaluated against the upward and downward changes observed at the tests level in one generation. Importantly, the computed invariance analyses support the conclusion that the constructs of interest tapped by the completed tests might not be meaningfully compared across cohorts and, therefore, only data at the test level should be considered when interpreting the pattern of changes. Nevertheless, the pattern observed for abstract reasoning (DAT-AR) and rote calculation (PMA-N) might support the statement that the fluid/crystallized distinction may be relevant here. In fact, RAPM and DAT-AR showed the highest correlation ($r = 0.47$) whereas RAPM and PMA-N showed a much lower correlation ($r = 0.17$).

5.2. Are global gains coming to an end in Spain? Is this the right question?

Flynn (2012, p. 83) stated that “an IQ score is not a number but a message.” The results reported in the present article suggest global intelligence gains are still active in Spain, although this estimate comes from upward and downward trends at the test level.

As described above, the 2022 cohort completed the Raven Advanced

Table 6
Estimated parameters for the partial scalar invariance bifactor model.

Model	Unstandardized			Standardized within-group	
	Loadings on <i>g</i>	Loadings on verbal	Intercepts	Loadings on <i>g</i>	Loadings on verbal
Quantitative Reasoning	0.80		0	0.79/0.79	
Vocabulary	0.37	0.36	-0.03	0.37/0.43	0.36/0.28
Abstract Reasoning	0.65/0.39		0/0.44	0.65/0.61	
Verbal Comprehension	0.29	0.55	0.02	0.29/0.35	0.54/0.46
Mental Rotation	0.37		0/-0.36	0.38/0.49	
Rote Calculation	0.52/0.24		0/-0.63	0.52/0.37	
Verbal Meanings	0.32	0.46	0/-0.63	0.32/0.33	0.46/0.33
Identical Figures	0.34		0/-0.33	0.33/0.39	
	<i>g</i>	Verbal			
Mean (DT)	0(1)/0.21(1.01)	0(1)/0.58(0.7)			

Note. “Unstandardized” parameters are standardized regarding the 1991 cohort (i.e., metric is such that all the variables are standardized in this cohort). When cells contain one value, it means that the parameter is constrained to be equal in both cohorts. If not, the cell shows 1991/2022 values.

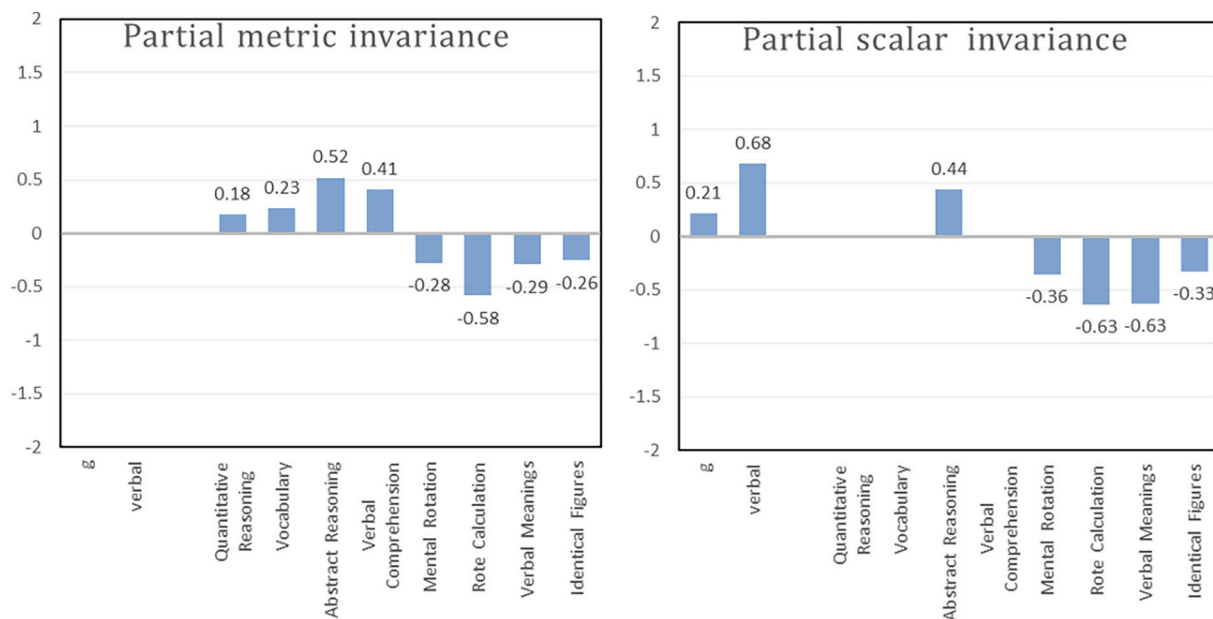


Fig. 2. Standardized differences (*d*) for factor mean differences, and intercept differences in each non-invariant test, depending on the model (difference was computed as cohort 2022 minus cohort 1991).

Progressive Matrices Test (RAPM) in addition to the same battery of cognitive tests completed by the 1991 cohort. We do have data from Spain obtained after administering the RAPM to university students since 1968 (Table 7). Colom et al. (1998) reported gains on the RAPM equivalent to 6.7 IQ points in one generation, from 1968 to 1996. The results observed twenty-six years later (present study, 2022) indicate that gains on the RAPM are probably over (Table 7).

We also do have data from Spain obtained after administering the abstract reasoning subtests from the DAT battery (DAT-AR) to university students since 1979. Colom et al. (1999) reported gains on the DAT-AR between 1979 and 1995 (16 years) equivalent to 5 IQ points (Table 7). If we compute the difference between the values obtained on this reasoning test by the 2022 cohort and the 1979 values on the same test (separated by 43 years) we get a *d* value of 1.0 which translates into a difference equivalent to 15 IQ points. As noted above, the correlation value between the RAPM and the DAT-AR for the 2022 cohort was $r = 0.47$. Although both mental tests presumably tap fluid/abstract reasoning, available results show no more generational gains on the RAPM but still remarkable gains on the DAT-AR, which reinforces the distinguishable nature of the changes observed at the test level.

What is then the right question?

Following Flynn (2018) we acknowledge that “society causes the development or atrophy of cognitive skills in terms of its own priorities

(p. 78).” The underlying mechanism might be quite like the explanation provided by Protzko (2015) for accounting for the fadeout effect that follows intervention programs aimed at increasing cognitive ability. The fadeout effect occurs because “children whose ability was increased lose their abilities once returned to their previous environment (...) there is no reason to believe that results from an intervention to raise intelligence should be permanent (...) the intelligence of children will react to the demands of the environments they are placed in, for good or for ill.” (p. 209).

This view is consistent with Flynn (2018) interpretation regarding the end of score gains on intelligence tests observed in some world regions: “although enriched environments dominated the 20th century, gains are not destined to persist like the law of gravity.” (p. 80). The candidate factors presumably contributing to these gains (increased schooling, demanding jobs, enhanced health and nutrition, reduced family size, heterosis, and the like) might begin to show diminishing returns. Quality of environments may cause downward and upward trends.

At the end of the day, the delineated general picture suggests that the messengers we use to evaluate cognitive ability changes across generations behave in a functionally independent way (Flynn, 2007). And this fact is consistent with the evidence that, within generations, these messengers show one remarkable dependence (correlate) which

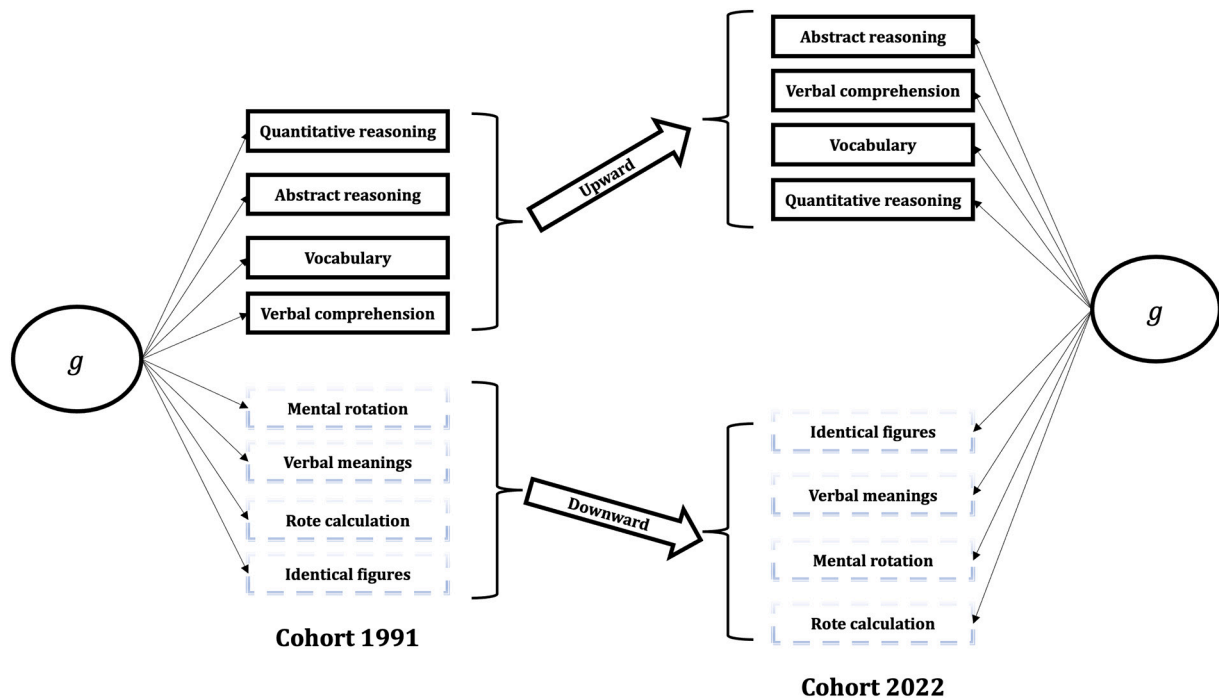


Fig. 3. Upward and downward changes on the eight intelligence tests completed by the 1991 and 2022 cohorts. The tests on the left are organized by their factor loadings (from highest to lowest, 0.73, 0.57, 0.49, 0.45, 0.41, 0.38, 0.38, and 0.30, respectively). The gains are observed on abstract reasoning, verbal comprehension, vocabulary, and quantitative reasoning [solid boxes], whereas losses are observed on rote calculation, mental rotation, verbal meanings, and identical figures [dotted boxes]. Although there is a global intelligence gain equivalent to 3.5 IQ points, these changes at the test level raise reservations regarding the appreciation of the general estimate. The largest blank space on the right between tests showing gains and losses helps to visualize the upward and downward trends at the test level.

Table 7

Generational changes on the RAPM and DAT-AR test in Spain. RAPM changes show values across 54 years (from 1968 to 2022) and DAT-AR changes show values across 43 years (from 1979 to 2022).

RAPM	N	Mean	SD
RAPM 1968 (Colom et al., 1998)	7335	20.94	6.19
RAPM 1996 (Colom et al., 1998)	3103	23.44	4.91
RAPM 2022 (Present Study)	349	23.04	4.25
DAT-AR			
DAT-AR 1979 (Colom et al., 1999)	440	32.8	8.4
DAT-AR 1995 (Colom et al., 1999)	1094	35.43	7.61
DAT-AR 2022 (Present Study)	349	40.50	4.4

strongly supports the positive manifold phenomenon documented across the planet (Warne & Burningham, 2019).

We now turn to the implications of this general picture for the available theories of the intelligence construct.

5.3. Implications for psychometric theories of intelligence

The analyses by Protzko and Colom (2021) of three classes of psychometric models (common cause, interconnected, and sampling models) regarding their position to accommodate brain lesion data led to the conclusion that common cause models did a better albeit not a perfect job. Specifically, bifactor models were able to accommodate the impact of local deficits along with implications derived from questionable concepts such as cognitive reserve. Sampling and interconnected

models showed weakness for accommodating the pattern observed in the lesion data.

Their discussion of the consequences for explanations of the positive manifold provided by these different psychometric models was based on the fact that focal chronic cortical lesions cause local instead of global or generalized effects. Although they were not looking for a winner, the key message was straightforward: psychometric models “are causal models, scientific theories, with necessarily causal connections and testable predictions (...) theories and explanations of the positive manifold must consider the fact that the covariance in the positive manifold must necessarily be causal in some capacity, yet manipulation of one local ability does not correspond to cross-sectional nor longitudinal effects on other abilities.”

From this perspective, the results reported here with respect to observed generational changes in a set of cognitive tests tapping reasoning, verbal ability, visuospatial ability, and quantitative ability, and implying positive (gains) and negative (losses) changes in one generation (30 years), suggest the presence of local instead of global effects. Although the positive manifold is evident within cohorts and shows one equivalent psychometric behavior, changes from the 1991 cohort to the 2022 cohort reveal positive and negative values without appreciable movements on the identified common higher-order factor within generations.

In this regard, Flynn (2012) acknowledged: “I never said real intelligence levels were raising.” (p. 138). It is unclear, however, how we should interpret this statement. If by ‘intelligence’ he meant the glue that binds all cognitive abilities together, then he is probably right. As discussed by Haier, Colom, and Hunt (2023) achieving the goal of enhancing the general ability devoted to make sense of all the specific abilities and skills, comprised by psychometric models such as the CHC (Schneider & McGrew, 2018) or the VPR (Johnson & Bouchard, 2005) models, might be doable hacking the brain and the genome. Nevertheless, if by ‘intelligence’ he meant the specific abilities and skills tapped by the messengers that scientists use to obtain the required information,

then he is plainly wrong.

These specific abilities and skills tapped by cognitive tests can go up and down without making any difference at the construct level. Thus, for instance, cognitive training programs can enhance some of these specific abilities without changing g (Anguera et al., 2013; Bejjanki, Zhang, Li, & Bavellier, 2014; Colom et al., 2013; Colom & Román, 2018). Also, the practice effect identified when solving cognitive tests and tasks improves the obtained scores but leave unchanged the tapped common latent factor (Estrada et al., 2015). Rodgers (1998) published a cogent commentary about the secular changes in intelligence test scores making a set of relevant questions. Question number 6 was this: “does the Flynn effect operate on all cognitive abilities, or only on certain abilities? Is the effect one that applies to latent intelligence itself, or to artifacts of the measurement of intelligence?” Beyond his answer to this question, he wrote: “one of the biggest puzzles growing out of the identification and research on the Flynn effect is the elusive status of causal explanations for the phenomenon.” (p. 342, emphasis added). Also, he underscored the idea that mean values and correlations tell different stories, something frequently forgotten (Hunt, 2011; Sauce & Matzel, 2018).

Gignac (2015) analyzed 85 years of data (from 1923 to 2008) obtained through the administration of the digit span subtest from the Wechslers. The main finding revealed not any increasing trend in either (digit span) forward and (digit span) backward recall. Given the documented very high relationship between the general factor of intelligence (g) and working memory capacity (Colom et al., 2005b, 2006, 2008; Oberauer et al., 2005), this finding may imply that the secular changes observed across several cognitive tests left unaffected across generations the higher-order ability that binds all abilities together (Haier et al., 2023).

If common cause models, especially bifactor models, can reasonably accommodate (a) the positive impact of intervention programs over specific cognitive abilities and skills, (b) the negative local instead of global impact of focal and chronic brain lesions, and (c) the simultaneous positive and negative changes in specific abilities and skills across generations, without commensurate changes in other correlated abilities and over the common higher-order factor, then perhaps we should consider seriously that they are more likely representations of the scientific concept of intelligence.

5.4. Limitations

The first limitation that must be raised is the non-representativeness of the considered cohorts. University freshmen comprised the 1991 and the 2022 cohorts and, therefore, the observed results may change if groups of, say, high school students were considered. Nevertheless, the cross-temporal meta-analysis of Raven's Progressive Matrices tests by Wongupparaj et al. (2015) spanning 64 years, from 1950 to 2014, concluded: “the Flynn effect is strong enough to be showed in even small and non-representative samples as well as in the very young and older adults.” (p. 8).

The second limitation is the lack of age and sex data for the 1991 cohort. For unknown reasons, the team of psychologists that administered the battery did not register these data. As shown in the participants' section, indirect sociodemographic data suggest both cohorts

might be considered roughly equivalent for the main purpose of the present study. Moreover, the pattern of results observed in our study for the eight cognitive ability tests completed by both cohorts provides evidence consistent with the conclusion that they may be reasonably equivalent. First, the correlation and factor matrices are strikingly similar. Second, we can disregard that the 1991 cohort shows higher (or lower) general learning ability than the 2022 cohort because of the identified upward and downward trends on the administered cognitive ability tests.

Finally, although the cognitive battery included eight heterogeneous standardized tests tapping reasoning, verbal ability, quantitative ability, and visuospatial ability, the resulting measurement model was far from ideal. It is highly recommended to obtain estimates at the ability level after the consideration of at least three measures/tests (Colom et al., 2013; Haier et al., 2023). However, the present study was conditioned by the battery administered to the first cohort and comparability between cohorts was a key and unnegotiable goal.

5.5. Conclusion

In conclusion, the findings reported here support Flynn (2007) statement that cognitive abilities do in fact show dependence within generations but can develop in a functionally independent way across generations. The invariance analyses reported here support this key point. The mechanism underlying the generational cognitive upward and downward changes observed in the present study might follow the unidirectional reactive model favored by Protzko (2015) when accounting for the fadeout effect identified after ending environmental intervention programs aimed at raising early intelligence. In a similar way, generational cognitive changes may be explained by sociological changes: “society causes the development or atrophy of cognitive skills in terms of its own priorities” (Flynn, 2018, p. 78). This one-way cultural perspective contradicts Lynn (2013) view favoring the nutrition theory of the secular increase in intelligence. He argued that most cultural hypotheses are falsified by the fact that the increases have been documented in infants. The cultural hypotheses “predict that the effect should be absent or minimal among infants and should increase progressively through childhood and adolescence as the environmental inputs have cumulative IQ boosting impacts” (Lynn, 2013, p. 768). However, if the nutrition theory is likely, then it is difficult to accommodate the fact that intelligence gains at the higher-order latent or construct level are not identified, unless we accept that the effect posited by the nutrition theory over the processing capacity of the brain is local instead of global and impacts on specific abilities, which is highly unlikely. Moreover, Pietschnig and Voracek (2015) meta-analytic findings revealed stronger generational gains for adults than for children.

Declaration of Competing Interest

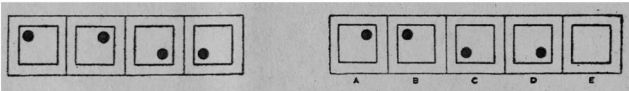
The authors declare no conflicts of interest.

Data availability

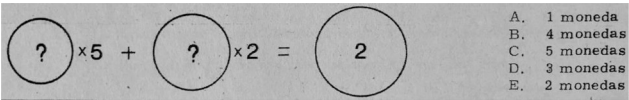
Data will be made available on request.

Appendix

Examples of items from the administered intelligence battery.
DAT-AR



Monedas-2



PMA-V

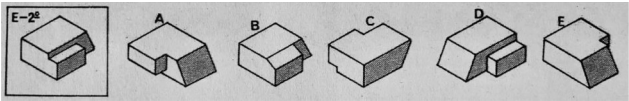
ANCIANO. (A) Seco, (B) Largo, (C) Feliz, (D) Viejo
Verbal Comprehension
A VECES UN GIGANTE NECESITA LOS SERVICIOS DE UN ENANO.

- A. El éxito es a menudo difícil de prever.
- B. A menudo se tiene necesidad de alguien inferior a uno.
- C. Dar es una cosa fácil.

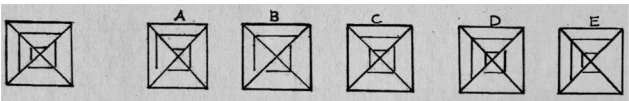
Verbal Meanings
Hizo un gesto de TEMOR.

(A) COMPASIÓN, (B) TRISTEZA, (C) MIEDO, (D) ASCO, (E) RISA

Rotation of Solid Figures



Identical Figures



PMA-N (Rote calculation)

(a)	(b)	(c)	
17	35	63	(a) - B M
84	28	17	(b) - B M
29	61	89	(c) - B M
140	124	169	

References

Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., Janowich, J., ... Gazzaley, A. (2013). Video game training enhances cognitive control in older adults. *Nature*, 501, 97–101.

Bejjanki, V. R., Zhang, R., Li, R., & Bavellier, D. (2014). Action videogame facilitates the development of better perceptual templates. *PNAS*, 111(47), 16961–16966.

Benson, A., Beaujean, A., & Taub, G. E. (2015). Using score equating and measurement invariance to examine the Flynn effect in the Wechsler adult intelligence scale. *Multivariate Behavioral Research*, 50(4), 398–415.

Bratsberg, B., & Rogeberg, O. (2018). Flynn effect and its reversal are both environmentally caused. *Proceedings of the National Academy of Sciences*, 115, 6674–6678.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233–255.

Colom, R., & Román, F. J. (2018). Enhancing intelligence: From the group to the individual. *Journal of Intelligence*, 6, 11. <https://doi.org/10.3390/jintelligence6010011>

Colom, R., et al. (1998). Generational IQ gains: Spanish data. *Personality and Individual Differences*, 25, 927–935.

Colom, R., et al. (1999). Are cognitive sex differences disappearing? Evidence from Spanish populations. *Personality and Individual Differences*, 27, 1189–1195.

- Colom, R., et al. (2005a). The generational intelligence gains are caused by decreasing variance in the lower half of the distribution: Supporting evidence for the nutrition hypothesis. *Intelligence*, 33, 83–91.
- Colom, R., et al. (2005b). Memory span and general intelligence: A latent-variable approach. *Intelligence*, 33, 623–642.
- Colom, R., et al. (2006). Complex span tasks, simple span tasks, and cognitive abilities: A reanalysis of key studies. *Memory & Cognition*, 34(1), 158–171.
- Colom, R., et al. (2008). Working memory and intelligence are highly related constructs, but why? *Intelligence*, 36, 584–606.
- Colom, R., et al. (2013). Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing. *Intelligence*, 41, 712–727.
- Dickens, W., & Flynn, J. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review*, 108(2), 346–369.
- Dutton, E., van der Linden, D., & Lynn, R. (2016). The negative Flynn effect: A systematic literature review. *Intelligence*, 59, 163–169.
- Eppig, C., Fincher, C. L., & Thornhill, R. (2010). Parasite prevalence and the worldwide distribution of cognitive ability. *Proceedings of the Royal Society B: Biological Sciences*, 277, 3801–3808.
- Estrada, E., et al. (2015). A general factor of intelligence fails to account for changes in tests' scores after cognitive practice: A longitudinal multi-group latent-variable study. *Intelligence*, 50, 93–99.
- Fabrigar, L. R., et al. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272–299.
- Flynn, J. R. (2007). *What is intelligence?* Cambridge University Press.
- Flynn, J. R. (2012). *Are we getting smarter? Rising IQ in the twenty-first century*. Cambridge University Press.
- Flynn, J. R. (2018). Reflections about intelligence over 40 years. *Intelligence*, 70, 73–83.
- Flynn, J. R., & Shayer, M. (2018). IQ decline and Piaget: Does the rot start at the top? *Intelligence*, 66, 112–121.
- Gignac, G. E. (2015). The magical numbers 7 and 4 are resistant to the Flynn effect: No evidence for increases in forward or backward recall across 85 years of data. *Intelligence*, 48, 85–95.
- Golino, H., & Christensen, A. P. (2022). EGAnet: Exploratory graph analysis—A framework for estimating the number of dimensions in multivariate data using network psychometrics. <https://CRAN.R-project.org/package=EGAnet>.
- Golino, H. F., & Demetriou, A. (2017). Estimating the dimensionality of intelligence like data using exploratory graph analysis. *Intelligence*, 62, 54–70.
- Gravetter, F. J., & Forzano, L. B. (2018). *Research methods for the behavioral sciences* (4th ed.). Wadsworth.
- Haier, R. J., Colom, R., & Hunt, E. B. (2023). *The science of human intelligence*. Cambridge University 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, 1–55.
- Hunt, E. B. (2011). *Human intelligence*. Cambridge University Press.
- Johnson, W., & Bouchard, T. (2005). The structure of human intelligence: It is verbal, perceptual, and imagen rotation (VPR), not fluid and crystallized. *Intelligence*, 33, 393–416.
- Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling. In R package version 0.5-6*. Retrieved from <https://CRAN.R-project.org/package=semTools>. Retrieved from.
- Kirby, K. N., & Gerlanc, D. (2013). BootES: An R package for bootstrap confidence intervals on effect sizes. *Behavior Research Methods*, 45, 905–927.
- Lim, S., & Jahng, S. (2019). Determining the number of factors using parallel analysis and its recent variants. *Psychological Methods*, 24(4), 452–467.
- Luong, R., & Flake, J. K. (2022). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychological Methods*. <https://doi.org/10.1037/met0000441>. Advance online publication.
- Lynn, R. (2009). What has caused the Flynn effect? Secular increases in the development quotient of infants. *Intelligence*, 37, 16–24.
- Lynn, R. (2013). Who discovered the Flynn effect? A review of early studies of the secular increase in intelligence. *Intelligence*, 41, 765–769.
- Manziona, J. M. (1978). *B.F.A.—Batería factorial de aptitudes [factorial abilities battery]*. Madrid: TEA, S.A.
- Mingroni, M. (2007). Resolving the IQ paradox: Heterosis as a cause of the Flynn effect and other trends. *Psychological Review*, 114, 806–829.
- Neisser, U. (1997). Rising scores on intelligence tests. *American Scientist*, 85, 440–447.
- te Nijenhuis, J., & Van Der Flier, H. (2013). Is the Flynn effect on g? A meta-analysis. *Intelligence*, 41, 802–807.
- Oberauer et al. (2005). Working memory and intelligence –their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, 131(1), 61–65.
- Pietschnig, J., & Voracek, M. (2015). One century of global IQ gains: A formal meta-analysis of the Flynn effect (1909–2013). *Perspectives on Psychological Science*, 10, 282–306.
- Protzko, J. (2015). The environment in raising early intelligence: A meta-analysis of the fadeout effect. *Intelligence*, 53, 202–210.
- Protzko, J. (2017). Effects of cognitive training on the structure of intelligence. *Psychonomic Bulletin & Review*, 24(4), 1022–1031.
- Protzko, J., & Colom, R. (2021). Testing the structure of human cognitive ability using evidence obtained from the impact of brain lesions over abilities. *Intelligence*, 89.
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71–90.
- R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Rodgers, J. L. (1998). A critique of the Flynn effect: Massive IQ gains, methodological artifacts, or both? *Intelligence*, 26(4), 337–356.
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.
- Roth, B., et al. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, 53, 118–137.
- Sauce, B., & Matzel, L. D. (2018). The paradox of intelligence: Heritability and malleability coexist in hidden gene-environment interplay. *Psychological Bulletin*, 144(1), 26–47.
- Schneider, W. J., & McGrew, K. S. (2018). The Cattell-horn-Carroll theory of cognitive abilities. In D. P. Flanagan, & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73–162). Guilford Press.
- Schooler, C. (1998). Environmental complexity and the Flynn effect. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 67–79). Washington, DC: American Psychological Association.
- Seisdedos, N. (1978). *Manual del test de Monedas [handbook of the Monedas test]*. Madrid: TEA, S.A.
- TEA, S.A. (1979a). *Manual del DAT (handbook of the DAT; 79 standardization)*. Madrid.
- TEA, S.A. (1979b). *Manual del PMA (handbook of the PMA; 79 standardization)*. Madrid.
- Teasdale, T. W., & Owen, D. R. (1987). National secular trends in intelligence and education: A twenty-year cross-sectional study. *Nature*, 325, 119–121.
- Trahan, L. H., Stuebing, K. K., Fletcher, J. M., & Hiscock, M. (2014). The Flynn effect: A meta-analysis. *Psychological Bulletin*, 140, 1332.
- Warne, R. W., & Burningham, C. (2019). Spearman's g found in 31 non-western nations: Strong evidence that g is a universal phenomenon. *Psychological Bulletin*, 145(3), 237–272.
- Wicherts, J. M., et al. (2004). Are intelligence test measurements invariant over time? Investigating the Flynn effect. *Intelligence*, 32, 509–538.
- Wongupparaj, P., Kumari, V., & Morris, R. G. (2015). A cross-temporal meta-analysis of Raven's progressive matrices: Age groups and developing versus developed countries. *Intelligence*, 49, 1–9.
- Woodley, M. A., Te Nijenhuis, J., Must, O., & Must, A. (2014). Controlling for increased guessing enhances the independence of the Flynn effect from g: The return of the brand effect. *Intelligence*, 43, 27–34.
- Yela, M. (1969). *Rotación de Figuras Macizas [rotation of solid figures]*. Madrid: TEA.
- Zajonc, R. B., & Mullanly, P. R. (1997). Birth order: Reconciling conflicting effects. *American Psychologist*, 52, 685–699.
- Hegelund, E. R., et al. (2021). The secular trend of intelligence test scores: The Danish experience for young men born between 1940 and 2000. *PLoS One*. <https://doi.org/10.1371/journal.pone.0261117>. December 9 2021.