Personality and Individual Differences 216 (2024) 112417

ELSEVIER

Personality and Individual Differences

Review

Is there a Flynn effect for attention? Cross-temporal meta-analytical evidence for better test performance (1990–2021)

Denise Andrzejewski [a,b,c,*], Elisabeth L. Zeilinger [d,e], Jakob Pietschnig [a]

[a] *Department of Developmental and Educational Psychology, Faculty of Psychology, University of Vienna, Austria*
[b] *Department of Psychology, School of Science and Technology, Middlesex University Dubai, United Arab Emirates*
[c] *Department of Psychology, School of Social Sciences, Heriot-Watt University Dubai, United Arab Emirates*
[d] *Academy for Ageing Research, Haus der Barmherzigkeit, Vienna, Austria*
[e] *Department of Clinical and Health Psychology, Faculty of Psychology, University of Vienna, Vienna, Austria*

A B S T R A C T

Generational IQ test score changes (i.e., the Flynn effect) have been observed for most measures of cognitive ability, although certain cognitive domains appear to be less affected by this effect than others. IQ test score changes have been found to differ between domains, but evidence of Flynn effects for specific IQ-related abilities is sparse. In the present cross-temporal meta-analysis, we investigate potential test score changes for attention as assessed by the d2 Test of attention. Based on data from 287 independent samples ($N = 21{,}291$) from 32 countries over a timespan of 31 years (1990–2021) we found evidence for moderate generational test score gains in concentration performance in adults, but not children.

* Corresponding author at: Department of Developmental and Educational Psychology, Faculty of Psychology, University of Vienna, Austria.
E-mail address: a12210477@unet.univie.ac.at (D. Andrzejewski).

https://doi.org/10.1016/j.paid.2023.112417
Received 14 July 2023; Received in revised form 6 September 2023; Accepted 11 September 2023
Available online 16 September 2023

investigation level may have masked so far only comparatively sparsely documented domain-specific changes. In fact, Rodgers (1998) pointed out that knowledge about domain-specific trajectories of generational IQ changes is critical to clarify the nature, meaning, and causes of the Flynn effect.

Observations of test score changes according to modern intelligence conceptualizations (e.g., the Cattell-Horn-Carroll-model of intelligence; Schneider & McGrew, 2013) that go beyond the mere distinction of crystallized and fluid intelligence indicate that IQ changes may be differentiated according to stratum II and consequently lower-echelon domains. A recent examination of stratum II abilities found Flynn effects for general domain-specific knowledge, learning-efficiency, and comprehension knowledge (Lazaridis et al., 2022). However, other stratum II domains showed a stagnation such as processing speed (Lazaridis et al., 2022). This finding conforms to earlier results which suggested no changes in inspection time (Nettelbeck & Wilson, 2004) and even negative changes for reaction time (Silverman, 2010).

Findings for the stratum II domains short-term memory and long-term storage and retrieval were less consistent. IQ gains have been observed for declarative memory (Rönnlund & Nilsson, 2008), visual learning and recall (Baxendale, 2010), as well as verbal and visuospatial short-term memory (Wongupparaj et al., 2017). However, stagnation (Gignac, 2015) and declines were reported for working memory capacity (Lazaridis et al., 2022), as well as negative Flynn effects for verbal and visuospatial working memory (Wongupparaj et al., 2017).

Within-domain trajectory differences may not necessarily be attributable to between-country differences alone but may be a consequence of cross-temporal variations in the Flynn effect strength and possibly direction. Such variations are to be expected because non-linear trajectories of the Flynn effect have been established globally (Pietschnig & Voracek, 2015) and within specific countries (Pietschnig & Gittler, 2015).

Investigating cross-temporal changes of intelligence and related domains is necessary for our understanding of the Flynn effect. Conceivably, IQ test score changes may be rooted in changes in executive functioning components, depending on a given domain. However, to date there is little evidence about such potential changes beyond some work regarding different memory domains (e.g., Baxendale, 2010). In particular, attention and attentional processes have been found to be associated with various measures of intelligence, including fluid intelligence, and working memory (Colom et al., 2008). While there is some evidence of a negative Flynn effect on the attention/working memory and learning trials of the CVLT-II/CVLT3 in the US (Graves et al., 2021), to date it is still unclear if there is a Flynn effect for selective attention.

Here, we address this gap in the literature by conducting a cross-temporal meta-analysis on the d2 Test of Attention (Brickenkamp, 1962), a well-established and widely-used measure of attention, from 1990 to 2021. To this end, we examine changes in concentration performance (i.e., reflecting selective and sustained attention) and test effectiveness (i.e., reflecting the relationship between processing speed and accuracy) over time. We further investigate influences of national macro indicators (i.e., GDP per capita, internet usage) on test performance in the respective time spans.

2. Methods

The present study was preregistered at the open science framework (doi:10.17605/OSF.IO/FCSD3). All deviations from the preregistration are documented at (https://osf.io/qu9gv).

2.1. Attention measure

We used the d2 Test of Attention (Brickenkamp, 1962), a well-established, validated (Dingel, 1971; Schmidt-Atzert & Ising, 1997), and widely-used measure to investigate cross-temporal changes in selective attention and processing speed. Originally, it has been developed for the use in Germanophone countries but has since been adapted and validated for the use in other countries (Bates & Lemay, 2004; Fernández-Marcos et al., 2018; Filippetti et al., 2022; Seisdedos & Brickenkamp, 2012). In validation studies, the d2 Test of Attention has shown excellent reliabilities with overall Cronbach alphas ranging between 0.90 and 0.96 (Brickenkamp, 1962; Brickenkamp et al., 2010).

The d2 Test of Attention is a cancellation test involving simultaneous presentation of visually similar stimuli and has been proposed to be a particularly useful measure of attention and concentration processes. The task is to cross out all target characters (a letter "d" with a total of two dashes placed above and/or below), which are interspersed with nontarget characters (a "d" with more or less than two dashes, and "p" characters with any number of dashes). Participants are given a time limit of 20 s to process each line and are typically unable to inspect all items in a line due to the time constraints. Traditionally the d2 Test of Attention was scored based on 14 lines with 47 characters each (658 total characters), however effective the release of the 10th edition only 12 lines are scored now (564 characters). The test can be administered both in individual or group-settings and typically takes between 8 and 10 min.

Test performance on the d2 Test of Attention is assessed in terms of two indices: Concentration Performance (CP) is measured by the number of correctly marked items (i.e., a measure for selective attention) minus the number of incorrectly identified items (i.e., commission errors) and Test Effectiveness (TNE) is measured by the number of all inspected items (i.e., a measure for processing speed) minus the number of incorrectly and non-identified items (i.e., commission plus omission errors). Concentration performance has been considered to be an indicator of selective and sustained attention. The TNE composite represents the relationship between speed and accuracy and therefore can serve as an indication of test taking style. There are further indices that can be interpreted for the d2 Test of attention. However, we presently focus on CP and TNE, because these indices represent the most informative commonly reported indices in the available literature.

2.2. Literature search

To identify articles reporting d2 Test of Attention scores, we first conducted a cited reference search for Brickenkamp (1962) and all subsequent d2 editions including non-Germanophone adaptations in ISI Web of Knowledge. This was deemed to be a reasonable strategy because any primary study that used the d2 Test should have cited the original test or any of its revisions. Moreover, we searched PsychInfo using the search strings "d2 AND attention" and "d2 Test of Attention" and screened the first 100 hits in www.scholar.google.com for the same search query. Finally, the open access theses and dissertations database (www.oatd.org) was searched for the terms "d2 Test of Attention" and "d2 Aufmerksamkeitstest".

Studies in languages other than English or German were translated with the free version of the online translator DeepL (https://deepl.com/translator; $k = 15$, 2, and 1 Spanish, Turkish, and French studies were included in the final sample). Our literature search was concluded in February 2023 (see Fig. 1 for the PRISMA study flowchart).

2.3. Inclusion and exclusion criteria

To be included in the current meta-analysis, studies had to meet the following inclusion criteria. Studies must have (i) had assessed attention via the pen-and-paper version of the d2 Test of Attention, (ii) reported mean scores and standard deviations on any of the d2 Test components (i.e., non-parametric reports were ineligible and reports that were based on unclear index calculations were excluded), (iii) investigated healthy participants, and (iv) been independent of other reported outcomes. In cases of data dependencies, the earliest reported results were preferred for study inclusion. In cases where essential information was missing, the corresponding authors of the respective studies were contacted with

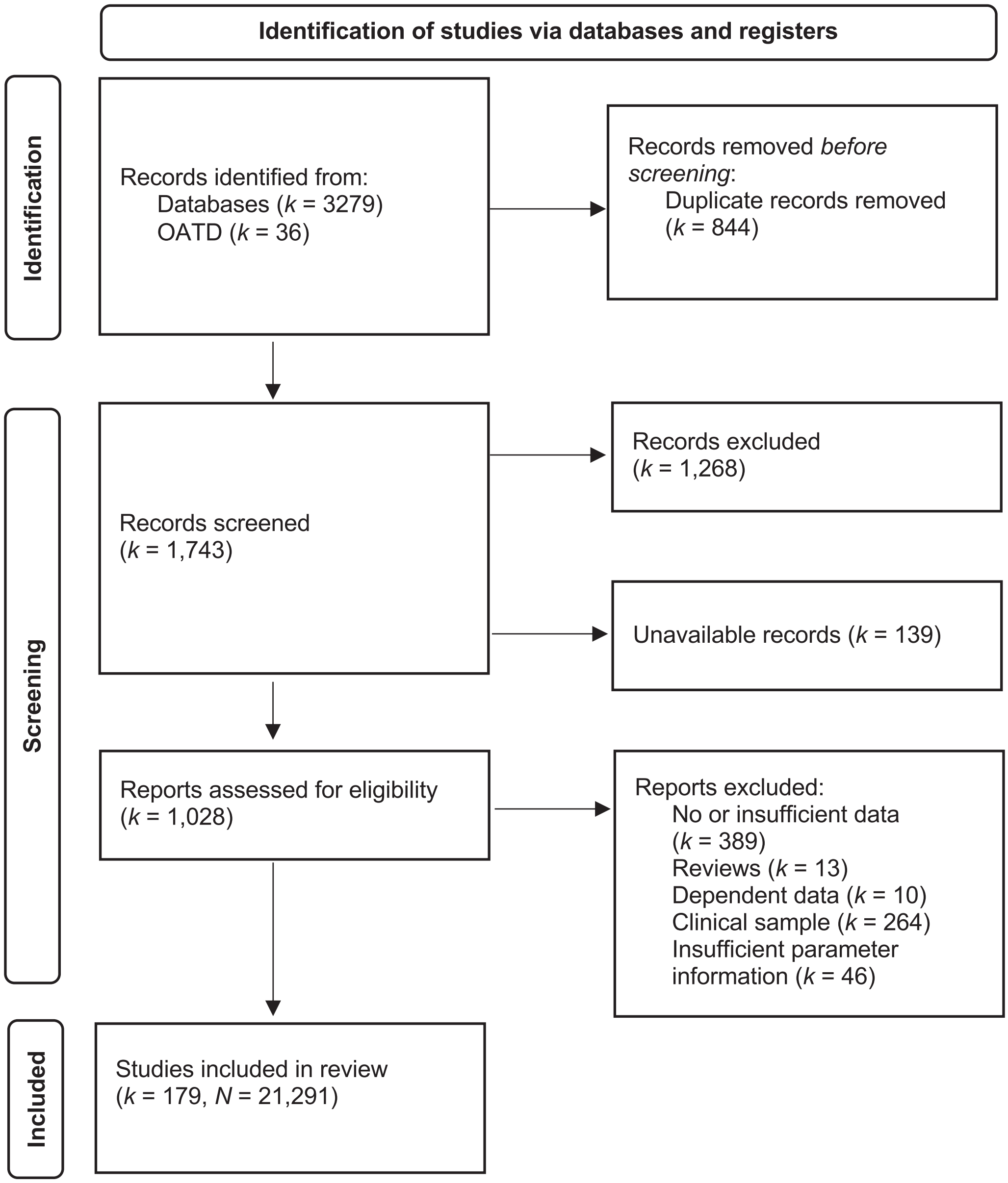


Fig. 1. Flowchart of study inclusion.

a follow up after two weeks. In cases where no responses were obtained, studies were excluded from analyses. In the current study we contacted in total 28 authors, three of which provided the requested missing information.

For eligible studies, essential details and characteristics were recorded including author(s) name(s), publication year, data collection year, country, study design (survey vs. experimental vs. clinical trial), data type (original vs. pre-test vs. control), sample size, sample type (children vs. adults), sample mean age and standard deviations, as well as the d2 Test edition used. Following standard approaches (Pietschnig & Gittler, 2015; Twenge & Campbell, 2001), we assumed that data collection took place two years prior to publication of a study if data collection years were not explicitly given. In cases where studies involved experimental manipulations, only pre-test data or data from control groups were included. We further recorded mean scores and standard deviations of d2 Test scores whenever these were provided.

In the final step we included annual country-specific macro-level indicators retrieved from the World Bank database (https://data.worldbank.org) as predictors. For prosperity we used GDP/PPP in USD (i.e., gross domestic product converted to international dollars using purchasing power parity rates). Everyday digital technology use was measured by internet use (% of population that indicated having used the internet in the past three months).

Data were coded twice independently by the first author [DA]. A reference list of included studies is provided in the online supplementary materials S1 (https://osf.io/qmb7j).

2.4. Final sample

Overall, we were able to retrieve data from 287 independent samples ($N = 21{,}291$) from 179 studies over a time span of 31 years (1990–2021). Samples originated from Germany (36.9 %, $k = 106$), Spain (13.6 %, $k = 39$), USA (6.6 %, $k = 19$) Austria (5.9 %, $k = 17$), and another 31 different countries (39.6 %, $k = 106$) with overall mean ages of 25.74 ($SD = 16.49$; 50.6 % women; see Table 1 for details).

Table 1
Sample characteristics.

	Concentration Performance (CP)			Total Test Effectiveness (TNE)		
		Mean (SD)	Min–Max		Mean (SD)	Min–Max
k	203	–	–	155	–	–
N	14,131	–	–	11,243	–	–
Sample size		69.61 (120.77)	6–1108		72.53 (95.41)	10–626
Mean score[a]		147.09	66–281		365.64	175–590
Year of data collection		–	1990–2021		–	1992–2020
Age of participant		25.21 (15.46)	7–72		23.82 (17.34)	7–74
Sex ratio (F/M)		50/50	0–100		50/48	0–100

	Concentration Performance (CP)		Total Test Effectiveness (TNE)	
	k	%	*k*	%
Country				
Germany	90	44.3	31	20
Spain	30	14.8	36	23.2
Austria	15	7.4	2	1.3
USA	9	4.3	15	9.7
Italy	8	3.8	9	5.8
Turkey	1	0.5	7	4.5
Canada	4	1.9	7	4.5
Argentina	1	0.5	7	4.5
Brazil	–	–	5	3.2
Denmark	2	1.0	4	2.6
Norway	1	0.5	4	2.6
India	4	2.0	4	2.6
Switzerland	4	2.0	2	1.8
Netherlands	8	3.8	1	0.6
Iran	1	0.5	1	0.6
Portugal	1	0.5	–	–
France	1	0.5	3	1.9
Poland	4	2.0	4	2.6
South Korea	3	1.5	1	0.6
New Zealand	4	2.0	1	0.6
Israel	3	1.5	3	1.9
Taiwan	1	0.5	2	1.3
Belgium	1	0.5	–	–
UK	1	0.5	1	0.6
Chile	4	2.0	4	2.6
Slovenia	1	0.5	–	–
Rep. Congo	1	0.5	–	–
Armenia	–	–	1	0.6

k = number of independent samples in which means were available for the analysis, *N* = sum of the sample size for each index.
[a] Weighted by sample size.

2.5. Statistical analyses

We used a Cross-Temporal Meta-Analysis approach to examine d2 Test score changes over time. Our main analyses focus on all included d2 Test data. We supplemented these by analyses from Germanophone samples only (i.e., the DACH region: Germany, Austria, and Switzerland). We interpret effect sizes according to the well-established criteria of Cohen (1988; η^2 = 0.01, 0.06, and 0.14 are interpreted as bottom thresholds of small, moderate, and large effects, respectively). Skewness and kurtosis of CP scores and TNE scores were acceptable yielding 0.190 and −0.151, −0.353 and −0.631, respectively. Data collection year exhibited a skewness of −1.690 and a kurtosis of 5.469.

In the d2 Test editions prior to 2010, scores were assessed based on 14 test lines with 47 characters each. However, in the subsequent revisions only 12 lines are used to calculate scores, resulting in a lower achievable raw score on the d2 Test. To ensure that results from different versions are cross-temporally comparable, we transformed raw scores reported in the primary studies into relative performance scores (i.e., the relative number of achievable scores instead of raw scores were calculated to account for the differences between maximum achievable scores between editions). By doing so, potential changes in overall test and concentration performance can consequently be interpreted as changes in population ability over time.

Following standard cross-temporal meta-analytical approaches (e.g. Pietschnig & Gittler, 2015; Twenge & Campbell, 2001), we predicted mean concentration and total test performance by year of data collection weighted by precision (i.e., sample size). We conducted the analyses separately for children and adults. Missing data in independent and dependent variables led to casewise exclusion from the respective analyses. Subsequently, the influence of national macro indicators on the indices were assessed. To this end, we ran multiple meta-regressions with data collection year and either prosperity (i.e. GDP/PPP in USD) or everyday digital technology use (i.e. internet use) as predictors. This approach was adopted because entering all predictors simultaneously would have resulted in considerable attrition due to data missingness (~75 % of cases would have had to be excluded).

Further, we used regression discontinuity analyses to examine potential changes in regression slopes attributed to the changes in the maximum achievable score of the revised d2 Test of Attention. Given the revision and re-standardization of the d2 Test in the 10th Edition, we reasoned that 2010 might be a point where discontinuities in trajectories may occur due to the revision and change in scoring (note, that by using relative test performance instead of raw scores, test performance remained comparable in our calculations). This approach enabled us to investigate potential influences of the revised test material.

We additionally conducted sensitivity and robustness analyses using

L1 regularized regressions (Tibshirani, 1996). In L1 regressions, the influences of potential leverage points are assessed by using absolute deviations of residuals from the regression line instead of squared deviations. We used the 'lmrob' function in the 'robustbase' R package to run the L1 regression analysis (Maechler et al., 2021).

To evaluate the overall trend of d2 Test scores over time, we computed annual mean scores weighted by sample size of the primary studies. This allowed us to describe changes in the trend of attention score data by means of joinpoint regression models (Kim et al., 2000). This approach allows the identification of points in a timeline where regression slopes change significantly in their strength. In other words, each joinpoint represents the point where two linear regression segments with significantly different slopes connect.

To determine the best-fitting model, we initially applied a simple model assuming linearity with no joinpoints. Subsequently, more complex models with additional joinpoints were fitted and compared to the simpler models using permutation tests that assess the squared errors of the null and alternative models. We employed Bonferroni-corrected permutation tests and used the Joinpoint Regression Program 4.0.4 (Statistical Research and Applications Branch, 2011) to fit models with up to four joinpoints. All data and the R analysis code are available at https://osf.io/utzmp and https://osf.io/xuy2e, respectively.

3. Results

3.1. Concentration Performance (CP)

For children, no meaningful time trends were identified in the meta-regression model for CP, thus indicating no changes between 2003 and 2020 in selective attention (see Fig. 2, Panel A; left side of Table 2). No meaningful effects of any of the macro indicators on test scores were observed (left side of Table 2).

Regression discontinuity analyses indicated that CP score trajectories did not differ significantly before and after 2010, thus indicating no effects of the test revision on CP changes ($b = 5.420$, $t = 1.107$, $p = .272$; see Supplemental Materials S2 at https://osf.io/5w6nk). Joinpoint regression did not show any significant changes in the slope of the observed trajectories, thus supporting linearity of changes (numerical details are provided in the online Supplemental Materials S3 at https://osf.io/dvuc2). Results of our precision-weighted L1 regressions were broadly in line with findings of our standard approaches. No significant time trends were observed in our single regression, ($b = -0.413$, $t = -0.352$, $p = .725$; see Supplemental Materials S4 at https://osf.io/u27ak). The precision-weighted L1 regression of data collection year and internet use yielded a positive effect ($b = 0.618$, $t = 2.661$, $p = .009$).

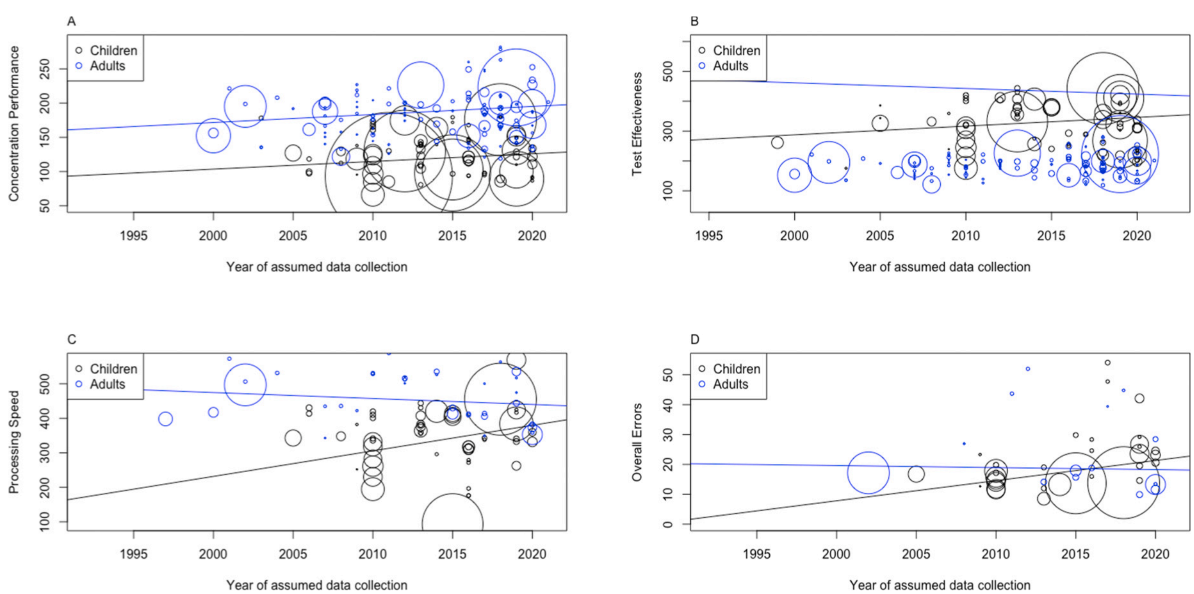
For adults, our analyses yielded a small significant positive effect of data collection year on CP scores ($b = 1.173$, $t = 2.708$, $p = .007$, $\eta_p^2 = 0.055$; right side of Table 2), indicating a meaningful increase in concentration performance between 2000 and 2021. When macro indicators were included, the effect of data collection year on test scores remained significant. Regression discontinuity analyses indicated that CP score trajectories did not differ significantly before and after 2010, thus indicating no effects of the test revision on CP changes ($b = -0.625$, $t = -0.263$, $p = .793$). No joinpoints were identified in joinpoint regressions indicating no meaningful changes in the CP change trajectory for adults over the investigated period. Results of our precision-weighted L1 regressions differed from the findings of our standard approaches. No significant time trends were observed in our single and multiple regressions. Furthermore, we observed a small positive effect for GDP ($b < 0.001$, $t = -2.140$, $p = .034$).

3.2. Test Effectiveness (TNE)

For children no meaningful time trends were identified in the meta-regression model for TNE, thus indicating no changes in test effectiveness ($b = 2.915$, $t = 1.467$, $p = .147$, $\eta_p^2 = 0.029$; Fig. 2, Panel B; left side of Table 3). No meaningful effects of any of the macro indicators on TNE scores were observed. Regression discontinuity and joinpoint regression indicated no significant changes in slopes and precision-weighted L1 regressions were in line with findings of our standard approaches.

For adults, our analyses yielded no significant effects of data collection year on TNE scores ($b = -1.904$, $t = -1.805$, $p = .074$, $\eta_p^2 = 0.039$; right side of Table 3), indicating no changes in test effectiveness. A small significant positive effect of data collection year was observed when GDP was added to the model ($b = -2.118$, $t = -1.998$, $p = .049$, $\eta_p^2 = 0.051$).

Regression discontinuity analyses and joinpoint regressions indicated no meaningful changes in the TNE change trajectory for adults over the investigated period. Results of our precision-weighted L1



Note. Symbol size is varied according to relative study weight within analysis

Fig. 2. Cross-temporal meta-regression of concentration performance, overall test effectiveness, processing speed, and overall errors and year of assumed data collection.

Table 2
Single and multiple weighted regression analyses for concentration performance for children and adults.

	Children							Adults						
	df	*(adj) R^2*	*b*	*SE*	*t*	*p*	η_p^2	*df*	*(adj) R^2*	*b*	*SE*	*t*	*p*	η_p^2
Single regression														
Year of assumed data collection	79	0.008	1.121	0.859	1.305	.196	0.021	125	0.047	1.173	0.433	2.708	.007	0.055
Multiple regressions														
Model 1														
Year of assumed data collection	76	0.013	1.441	0.915	1.573	.120	0.031	118	0.028	0.890	0.430	2.067	.041	0.035
GDP			<0.001	<0.001	−1.139	.258	0.016			<0.001	<0.001	−0.663	.508	0.003
Model 2														
Year of assumed data collection	76	0.022	0.718	0.939	0.765	.447	0.007	118	0.032	1.186	0.487	2.435	.016	0.047
Internet use			0.351	0.249	1.411	.162	0.025			−0.917	0.982	−0.967	.335	0.007

Note. Cases were precision-weighted according to sample size.

Table 3
Single and multiple weighted regression analyses for test effectiveness for children and adults.

	Children							Adults						
	df	*(adj) R^2*	*b*	*SE*	*t*	*p*	η_p^2	*df*	*(adj) R^2*	*b*	*SE*	*t*	*p*	η_p^2
Single regression														
Year of assumed data collection	72	0.015	2.915	1.987	1.467	.147	0.029	79	0.027	−1.904	1.055	−1.805	.074	0.039
Multiple regressions														
Model 1														
Year of assumed data collection	69	−0.007	2.504	2.288	1.094	.278	0.017	73	0.043	−2.118	1.060	−1.998	.049	0.051
GDP			<0.001	<0.001	0.337	.737	0.002			<0.001	<0.001	−0.966	.322	0.013
Model 2														
Year of assumed data collection	69	0.003	0.764	3.018	0.253	.801	<0.001	73	0.030	−2.162	1.714	−1.261	.211	0.021
Internet use			0.839	0.906	0.927	.357	0.012			−0.017	0.470	−0.038	.970	<0.001

Note. Cases were precision-weighted according to sample size.

regressions were mostly in line with the findings of our standard approaches. When macro indicators were added to our single regression models, effects of data collection year on test scores showed a positive effect for GDP ($b < 0.001$, $t = -2.140$, $p = .034$), but no effect for internet use.

3.3. Processing speed

Data collection year had a moderate positive, albeit nominally non-significant, cross-temporal effect ($b = 7.398$, $t = 1.920$, $p = .06$, $\eta_p^2 = 0.065$) on processing speed in children (left side of Table 4; Fig. 2 Panel C). Moderate significant positive effects of data collection year were observed when GDP ($b = 8.406$, $t = 2.039$, $p = .046$, $\eta_p^2 = 0.075$) and internet use ($b = 10.446$, $t = 2.052$, $p = .045$, $\eta_p^2 = 0.076$) were added to the model. Regression discontinuity and joinpoint regressions did not indicate any significant changes in regression slopes. Precision-weighted L1 regressions showed no significant influences of data collection year when macro indicators were added to the models.

For adults, our analyses yielded no significant effects of data collection year on processing speed ($b = -1.741$, $t = -1.329$, $p = .192$, $\eta_p^2 = 0.045$). No significant effects of any of the macro indicators on test scores were observed. Regression discontinuity analyses and joinpoint regressions did not indicate any change in regression slopes. Precision-weighted L1 regressions showed a significant negative effect for internet use on processing speed ($b = -1.329$, $t = -4.504$, $p < .001$).

3.4. Overall errors

For children, our analyses yielded a significant large positive effect of data collection year on the number of committed errors ($b = 0.675$, $t = 2.163$, $p = .038$, $\eta_p^2 = 0.131$), indicating a meaningful increase in overall errors between 2000 and 2021 (Table 5; Fig. 2 Panel D). When macro indicators were added to the models, effect of data collection year on test scores remained significant, excepting when GDP was added as a

Table 4
Single and multiple weighted regression analyses for processing speed for children and adults.

	Children							Adults						
	df	*(adj) R^2*	*b*	*SE*	*t*	*p*	η_p^2	*df*	*(adj) R^2*	*b*	*SE*	*t*	*p*	η_p^2
Single regression														
Year of assumed data collection	53	0.047	7.398	3.854	1.920	.060	0.065	37	0.019	−1.741	1.310	−1.329	.192	0.045
Multiple regressions														
Model 1														
Year of assumed data collection	51	0.039	8.405	4.122	2.039	.046	0.075	32	0.073	−2.387	1.332	−1.792	.082	0.091
GDP			>−0.001	<0.001	−0.698	.488	0.009			>−0.001	<0.001	−0.895	.377	0.024
Model 3														
Year of assumed data collection	51	0.045	10.446	5.090	2.052	.045	0.076	32	0.058	−1.782	2.014	−0.885	.383	0.023
Internet use			−1.279	1.423	−0.899	.372	0.015			−0.320	0.604	−0.530	.599	0.008

Note. Cases were precision-weighted according to sample size.

Table 5
Single and multiple weighted regression analyses for overall errors for children and adults.

	Children							Adults						
	df	*(adj)* R^2	*b*	*SE*	*t*	*p*	η_p^2	*df*	*(adj)* R^2	*b*	*SE*	*t*	*p*	η_p^2
Single regression														
Year of assumed data collection	31	0.103	0.675	0.312	2.163	.038	0.131	13	−0.073	−0.068	0.338	−0.203	.842	0.003
Multiple regressions														
Model 1														
Year of assumed data collection	29	0.077	0.696	0.348	1.997	.055	0.121	10	−0.011	−0.008	3.212	−0.025	.980	<0.001
GDP			>−0.001	<0.001	−0.028	.978	<0.001			>−0.001	<0.001	−1.353	.206	0.154
Model 2														
Year of assumed data collection	29	0.173	1.632	0.595	2.742	.010	0.205	10	−0.061	0.372	0.503	0.740	.476	0.051
Internet use			0.104	0.201	−1.839	.076	0.104			−0.262	0.233	−1.126	.286	0.112

Note. Cases were precision-weighted according to sample size.

predictor to the model ($b = 0.696$, $t = 1.997$, $p = .055$, $\eta_p^2 = 0.121$). Regression discontinuity and joinpoint regressions indicated no significant slope changes. Precision-weighted L1 regressions yielded no significant effects of data collection year or macro indicators on test performance.

For adults, our analyses yielded no significant effect of data collection year on overall errors ($b = -0.068$, $t = -0.203$, $p = .842$, $\eta_p^2 = 0.003$). No meaningful effects of any of the macro indicators on test scores were observed. Furthermore, regression discontinuity and joinpoint regressions indicated no significant slope changes. Similarly, precision-weighted L1 regressions yielded no significant influences of data collection year or macro-indicators, excepting a significant negative effect of internet use ($b = -0.427$, $t = -5.009$, $p < .001$).

3.5. Germanophone samples

To assess robustness of our time trends, we repeated all our analyses in a data subset of observations from Germanophone countries (i.e., the DACH region). Our results were broadly in line with our main analysis (see Supplementary Materials S5 at https://osf.io/fr2mu). Once again, for children no meaningful time trends for CP were observed, nor did macro-level indicators yield significant effects. The time trends for CP in adults found in the main analysis were not observed in the DACH region ($b = 0.615$, $t = 1.156$, $p = .251$, $\eta_p^2 = 0.015$). For TNE the results were virtually identical with our main analysis. However, the time trend for overall errors in children in our main analysis did not replicate for children in the DACH region ($b = 0.090$, $t = 0.241$, $p = .815$, $\eta_p^2 = 0.006$), nor were there significant effects for macro-indicators observed. Interestingly, in contrast to our main analysis findings, data collection year showed a significant negative effect for overall errors in children ($b = -5.262$, $t = 1.445$, $p = .006$).

Regression discontinuity and joinpoint regression analyses yielded no significant slope changes. Results of our precision-weighted L1 regressions yielded a significant negative effect of data collection year on TNE for adults, indicating a decline of test effectiveness in adults in the DACH region ($b = -4.403$, $t = 1.251$, $p < .001$). Supplementary analyses for cross-temporal regressions in data subsets of developed and developing countries (United Nations, 2014) can be found in the online Supplementary Table S7.

4. Discussion

Here, we investigated generational test score changes in attention in a large number of incidentally recruited samples. Specifically, we examined changes in concentration performance (i.e., selective and sustained attention) and test effectiveness (i.e., speed and accuracy) on the d2 Test of Attention in adults and children between 1990 and 2021. Our results indicate that concentration performance in adults has been increasing over the investigated time period. Moreover, our analyses yielded evidence for increasing overall errors and somewhat increased processing speed in children, indicating a shift in test taking styles. These results present several points of interest as we discuss below.

4.1. A Flynn effect for attention

A clear positive meaningful Flynn effect was observed for concentration performance in adults. This mirrors findings of generational gains found in other executive functions, such as working memory (Baxendale, 2010; Wongupparaj et al., 2017). Attention and working memory have previously been shown to be associated with fluid intelligence (Colom et al., 2008). Given this link, it is possible that the observed fluid IQ gains are rooted in changes in these executive function components.

Some non-significant, but meaningful gains in concentration performance were observed for children. Although the failure to reach nominal statistical significance for children may be due to a comparatively lower study power, we observed in any case stronger gains for adults than children. This is consistent with previous reports of stronger IQ gains of adults than children in standard intelligence testing (Flynn, 2010; Pietschnig et al., 2013). It has been argued that better education could explain age differentiated Flynn effects (Flynn, 2010) which could conceivably explain the presently observed pattern.

Our observed change trajectories indicated stronger changes in all data than when we focused on Germanophone samples only. This observation can be attributed to two potential causes. On the one hand, the larger strength of the observed global change trajectory may have been a consequence of the admixture of country-specific change trajectories. This interpretation is arguably in line with country-specific patterns that have been observed for the Flynn effect in the past (Pietschnig & Voracek, 2015). On the other hand, isolated data points from individual countries may have acted as leverage points in our regression analysis, thus arbitrarily inflating results of global ordinary least squares regressions. This interpretation is supported by lower gains for global data in our mean absolute deviation regression analyses. However, Germanophone gains in concentration performance of adults remained meaningful, thus corroborating the salience of this effect.

National economic prosperity appeared to only play a minor role for our time trends. Specifically, in our analyses GDP showed mostly trivial and small positive effects for adults and children, respectively. This somewhat contrasts previous findings of positive associations between national wealth and the Flynn effect in several nations (Lynn & Vanhanen, 2002; Pietschnig & Voracek, 2015) and indicates that the role of GDP for changes in attention is small at best.

Internet use predicted concentration performance positively, yielding small effects for children but no meaningful effects for adults. This seems to be in contrast with findings that indicate adverse effects of digitalization in general, and video games, media multitasking, as well as overall increased screen time on attention capabilities in particular (Swing et al., 2010; for a review see Vedechkina & Borgonovi, 2021; Zheng et al., 2014). However, modern technology exposure has been conceptually linked to the Flynn effect (Clark et al., 2016; Neisser, 1997) which is in line with our observation of positive associations between internet use and attention. However, this association was small in terms of strength which is unsurprising, given previous evidence that has indicated a limited relevance of technology for the Flynn effect (Dutton et al., 2016; Flynn, 2012; Pietschnig, 2016).

These results further support the importance to understand the Flynn effect as a domain-specific phenomenon, that must be assumed to be differentiated according to stratum II IQ and its related domains (e.g., stratum II abilities: Lazaridis et al., 2022; spatial ability: Pietschnig & Gittler, 2015, emotional intelligence: Pietschnig & Gittler, 2017; working memory: Wongupparaj et al., 2017).

4.2. Test taking styles: test effectiveness, processing speed, and errors

Changes in test effectiveness are a composite measure of processing speed and errors, thus arguably representing changes in test taking styles. Our results indicate differentiated trajectories of test taking styles for children and adults. Specifically, we observed a small increase in test effectiveness for children accompanied by increased processing speed and a moderate-to-large significant increase in overall errors. In contrast, for adults our results showed small declines in test effectiveness, processing speed, and overall errors. A faster and less accurate test taking style, as seen here in children, is considered to represent impulsivity, while a more accurate but slower performance represents reflexivity (Brickenkamp et al., 2010; Kagan, 1966). Thus, our findings indicate that children have become increasingly more impulsive in their test taking styles in the past three decades. Conceivably, this may be attributed to effects of an increasingly permissive society which has been suggested to be less likely to sanction errors (e.g., Brand, 1990).

The moderate increase in processing speed for children contrasts prior Flynn effect research on inspection time in Australian children where no meaningful changes were observed (Nettelbeck & Wilson, 2004). However, these findings should be taken with a grain of salt, because reanalyses of this trajectory in Germanophone samples only showed a reversal of the observed sign. This raises concerns about the stability of these observed changes.

However, trajectories of adult samples showed a small decline in processing speed in all data as well as in Germanophone samples. This observation contrasts evidence from previous studies that investigated test taking style-related changes in adults and showed no changes in test taking speed (Must & Must, 2018) in Estonia or speed-related stratum II abilities (Gs and Gt) in Germanophone samples (Lazaridis et al., 2022).

Our results for overall errors were once again differentiated according to participant age. We observed a moderate-to-large significant increase in overall errors in children, but no meaningful change for adults. The only meaningful macro-level driver of errors appeared to be a positive moderate-to-large effect of internet use on errors in children. Neither GDP in either age group nor internet use in adults showed any meaningful associations. This is interesting, because it contrasts our observations of positive effects of internet use on concentration performance. This observation is inconsistent with the proposed positive effects of modern technology on the Flynn effect (e.g., Neisser, 1997) and further points towards the limited relevance of technology of test score changes (Pietschnig, 2016).

These findings support the interpretation of an increase in impulsive test taking styles in children which manifests in a significant increase in errors and processing speed. There are various potential reasons for this observation. First, instructions provided during tests that emphasize speed rather than accuracy may contribute to this behavior. When individuals are instructed to complete tasks quickly, they may prioritize speed over accuracy, leading to more impulsive responses (Ackerman & Ellingsen, 2016).

Second, beliefs about speed being equated with success or competence (Wieselmann et al., 2020) may have resulted in increasingly impulsive test-taking styles. The erroneously held perception that faster performance is indicative of higher ability may once more encourage individuals to prioritize speed at the expense of accuracy (e.g., participants may erroneously expect to be rewarded for a speedy performance).

Third, a perceived lack of consequences in regard to the test outcome may lead to a low motivation to cognitively invest more than surface level processing in test responses (i.e., akin to the competency-performance problem), thus leading to a more impulsive as opposed to a reflective test taking style and consequently more errors (Jonassen & Grabowski, 1993, p. 113). This idea is rooted in the assumption that individuals may be more likely to invest effort and exhibit focused attention when there are tangible consequences of their actions, such as grades or rewards that are at stake (Frömer et al., 2021). Particularly in cross-temporally increasingly permissive societies (Brand, 1990), such effects could be expected to become more pronounced.

4.3. Limitations

First, the samples that have been included in this meta-analysis were heterogeneous in terms of nationality which conceivably may have led to an apples-and-oranges problem. However, we conducted robustness analyses for our results by rerunning all our analyses based on Germanophone samples (i.e., a rather homogeneous data subset in terms of nationality) and observed broadly similar results. In a similar vein, not all studies included in the current cross-temporal meta-analysis are population representative which arguably may have introduced statistical noise into our data. This means that existing true effects become harder to detect. Therefore, our identified effects should be considered to represent conservative estimates of the true effects.

Second, information about the used test edition within studies were often missing. This means, that we were unable to determine in many cases which one of the two existing administration modes (i.e., original vs. revised procedure) was used. However, we used discontinuity regressions and joinpoint regressions, to examine cross-temporal changes in the trajectory slopes which showed no indication for any systematic influences of potentially unidentified test edition-related changes.

Finally, it needs to be acknowledged that the national macro-level indicators that we used in our moderator analyses represent rather crude proxies for potential influential variables. For instance, data on internet use provides information on the percentage of the population that has used the internet in the past three months. While this information offers insight into the adoption of internet technology at a broad level, it does not provide sufficient granularity to assess the magnitude or intensity of internet use among the individuals within the respective samples. This means that the presently observed effects of macro-indicators must be considered to represent a lower threshold of the true effect due to the considerable statistical noise that is introduced by such crude measures.

5. Concluding remarks

In all, we show here first evidence for a Flynn effect for attention. This effect appears to be differentiated according to participant age, indicating moderate positive concentration performance changes for adults, but not for children. Conceivably, this may indicate a meaningful role of executive functions for changes in more traditional IQ domains. Cross-temporally increasing error rates and processing speed of children may be attributed to more impulsive behaviors whilst taking tests.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.paid.2023.112417.

CRediT authorship contribution statement

Denise Andrzejewski: Formal analysis, Visualization, Writing-original draft preparation; Elisabeth L. Zeilinger: Writing-review & editing; Jakob Pietschnig: Conceptualization, Methodology, Writing-review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. No funding has been recieved for this paper.

Data availability

We have shared the data on OSF and link is available in manuscript

References

Ackerman, P. L., & Ellingsen, V. J. (2016). Speed and accuracy indicators of test performance under different instructional conditions: Intelligence correlates. *Intelligence, 56*, 1–9. https://doi.org/10.1016/j.intell.2016.02.004

Bates, M., & Lemay, E. (2004). The d2 test of attention: Construct validity and extensions in scoring techniques. *Journal of the International Neuropsychological Society, 10*(3), 392–400. https://doi.org/10.1017/S135561770410307X

Baxendale, S. (2010). The Flynn effect and memory function. *Journal of Clinical and Experimental Neuropsychology, 32*(7), 699–703. https://doi.org/10.1080/13803390903493515

Brand, C. (1990). A 'gross' underestimate of a 'massive' IQ rise? A rejoinder to Flynn. *The Irish Journal of Psychology, 11*(1), 52–56. https://doi.org/10.1080/03033910.1990.10557788

Bratsberg, B., & Rogeberg, O. (2018). Flynn effect and its reversal are both environmentally caused. *Proceedings of the National Academy of Sciences, 115*(26), 6674–6678. https://doi.org/10.1073/pnas.1718793115

Brickenkamp, R. (1962). *Aufmerksamkeits-Belastungs-Test Handanweisung d-2.* Hogrefe.

Brickenkamp, R., Liepmann, D., & Schmidt-Atzert, L. (2010). *Test d2 - Revision: Aufmerksamkeits- und Konzentrationstest.* Hogrefe.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1–22. https://doi.org/10.1037/h0046743

Clark, C. M., Lawlor-Savage, L., & Goghari, V. M. (2016). The Flynn effect: A quantitative commentary on modernity and human intelligence. *Measurement: Interdisciplinary Research and Perspectives, 14*(2), 39–53. https://doi.org/10.1080/15366367.2016.1156910

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Colom, R., Abad, F. J., Quiroga, M.Á., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs, but why? *Intelligence, 36*(6), 584–606. https://doi.org/10.1016/j.intell.2008.01.002

Dingel, W. (1971). Mitteilung über Zuverlässigkeitsaspekte des Test d2. *Diagnostica, 17*, 137–140.

Dutton, E., & Lynn, R. (2013). A negative Flynn effect in Finland, 1997–2009. *Intelligence, 41*(6), 817–820. https://doi.org/10.1016/j.intell.2013.05.008

Dutton, E., & Lynn, R. (2015). A negative Flynn effect in France, 1999 to 2008–9. *Intelligence, 51*, 67–70. https://doi.org/10.1016/j.intell.2015.05.005

Dutton, E., van der Linden, D., & Lynn, R. (2016). The negative Flynn effect: A systematic literature review. *Intelligence, 59*, 163–169. https://doi.org/10.1016/j.intell.2016.10.002

Dworak, E. M., Revelle, W., & Condon, D. M. (2023). Looking for Flynn effects in a recent online U.S. adult sample: Examining shifts within the SAPA Project. *Intelligence, 98*, 101734. https://doi.org/10.1016/j.intell.2023.101734

Fernández-Marcos, T., De La Fuente, C., & Santacreu, J. (2018). Test–retest reliability and convergent validity of attention measures. *Applied Neuropsychology: Adult, 25*(5), 464–472. https://doi.org/10.1080/23279095.2017.1329145

Filippetti, V., Gutierrez, M., Krumm, G., & Mateos, D. (2022). Convergent validity, academic correlates and age- and SES-based normative data for the d2 test of attention in children. *Applied Neuropsychology-Child.* https://doi.org/10.1080/21622965.2021.1923494

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*(1), 29–51. https://doi.org/10.1037/0033-2909.95.1.29

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*(2), 171–191. https://doi.org/10.1037/0033-2909.101.2.171

Flynn, J. R. (2010). Problems with IQ gains: The huge vocabulary gap. *Journal of Psychoeducational Assessment, 28*(5), 412–433. https://doi.org/10.1177/0734282910373342

Flynn, J. R. (2012). Are we getting smarter? Rising IQ in the twenty-first century. Cambridge University Press.

Frömer, R., Lin, H., Dean Wolf, C. K., Inzlicht, M., & Shenhav, A. (2021). Expectations of reward and efficacy guide cognitive control allocation. *Nature Communications, 12*(1), 1030. https://doi.org/10.1038/s41467-021-21315-z

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. https://doi.org/10.1016/j.intell.2014.11.00

Gonthier, C., & Grégoire, J. (2022). Flynn effects are biased by differential item functioning over time: A test using overlapping items in Wechsler scales. *Intelligence, 95*, 101688. https://doi.org/10.1016/j.intell.2022.101688

Graves, L. V., Drozdick, L., Courville, T., Farrer, T. J., Gilbert, P. E., & Delis, D. C. (2021). Cohort differences on the CVLT-II and CVLT3: Evidence of a negative Flynn effect on the attention/working memory and learning trials. *The Clinical Neuropsychologist, 35*(3), 615–632. https://doi.org/10.1080/13854046.2019.1699605

Jonassen, D. H., & Grabowski, B. L. H. (1993). *Handbook of individual differences, learning, and instruction.* L. Erlbaum Associates.

Kagan, J. (1966). Reflection-impulsivity: The generality and dynamics of conceptual tempo. *Journal of Abnormal Psychology, 71*(1), 17–24. https://doi.org/10.1037/h0022886

Kim, H. J., Fay, M. P., Feuer, E. J., & Midthune, D. N. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine, 19*(3), 335–351. https://doi.org/10.1002/(sici)1097-0258(20000215)19:3<335::aid-sim336>3.0.co;2-z

Lazaridis, A., Vetter, M., & Pietschnig, J. (2022). Domain-specificity of Flynn effects in the CHC-model: Stratum II test score changes in Germanophone samples (1996–2018). *Intelligence, 95*, Article 101707. https://doi.org/10.1016/j.intell.2022.101707

Lynn, R., & Vanhanen, T. (2002). *IQ and the Wealth of Nations.* Praeger.

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., … Conceicao, E. L. T. (2021). *Robustbase: Basic Robust Statistics [Computer software]*. & Anna di Palma, M http://robustbase.r-forge.r-project.org/.

Must, O., & Must, A. (2018). Speed and the Flynn effect. *Intelligence, 68*, 37–47. https://doi.org/10.1016/j.intell.2018.03.001

Must, O., Must, A., & Raudik, V. (2003). The secular rise in IQs: In Estonia, the Flynn effect is not a Jensen effect. *Intelligence, 31*(5), 461–471. https://doi.org/10.1016/S0160-2896(03)00013-8

Neisser, U. (1997). Rising Scores on Intelligence Tests: Test scores are certainly going up all over the world, but whether intelligence itself has risen remains controversial. *American Scientist, 85*(5), 440–447.

Nettelbeck, T., & Wilson, C. (2004). The Flynn effect: Smarter not faster. *Intelligence, 32*(1), 85–93. https://doi.org/10.1016/S0160-2896(03)00060-6

Pietschnig, J. (2016). The Flynn effect: Technology may be part of it, but is most certainly not all of it. *Measurement: Interdisciplinary Research and Perspectives, 14*(2), 70–73. https://doi.org/10.1080/15366367.2016.1171612

Pietschnig, J., & Gittler, G. (2015). A reversal of the Flynn effect for spatial perception in German-speaking countries: Evidence from a cross-temporal IRT-based meta-analysis (1977–2014). *Intelligence, 53*, 145–153. https://doi.org/10.1016/j.intell.2015.10.004

Pietschnig, J., & Gittler, G. (2017). Is ability-based emotional intelligence impervious to the Flynn effect? A cross-temporal meta-analysis (2001–2015). *Intelligence, 61*, 37–45. https://doi.org/10.1016/j.intell.2016.12.006

Pietschnig, J., Tran, U. S., & Voracek, M. (2013). Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypothesis perhaps. *Intelligence, 41*(6), 791–801. https://doi.org/10.1016/j.intell.2013.06.005

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*(3), 282–306. https://doi.org/10.1177/1745691615577701

Rodgers, J. L. (1998). A critique of the Flynn Effect: Massive IQ gains, methodological artifacts, or both? *Intelligence, 26*(4), 337–356. https://doi.org/10.1016/S0160-2896(99)00004-5

Rönnlund, M., & Nilsson, L.-G. (2008). The magnitude, generality, and determinants of Flynn effects on forms of declarative memory and visuospatial ability: Time-sequential analyses of data from a Swedish cohort study. *Intelligence, 36*(3), 192–209. https://doi.org/10.1016/j.intell.2007.05.002

Rushton, J. P. (1999). Secular gains in IQ not related to the g factor and inbreeding depression — Unlike black-white differences: A reply to Flynn. Undefined. https://www.semanticscholar.org/paper/Secular-gains-in-IQ-not-related-to-the-g-factor-and-Rushton/e2bc79c4c5b0b7187d5c38ab51a65f1676a6f724.

Schmidt-Atzert, L., & Ising, M. (1997). Ein Beitrag zur Konstruktvalidität von d2 und Revisionstest. *Zeitschrift Für Differentielle und Diagnostische Psychologie, 18*(4), 241–250.

Schneider, J., & McGrew, K. (2013). *The Cattell-Horn-Carroll (CHC) Model of Intelligence v2.2: A visual tour and summary*. http://www.iapsych.com/chcv2.pdf.

Seisdedos, C. N., & Brickenkamp, R. (2012). Test de atencion d2. In *TEA Ediciones Madrid*. https://biblioteca.uazuay.edu.ec/buscar/item/77000.

Silverman, I. W. (2010). Simple reaction time: It is not what it used to be. *The American Journal of Psychology, 123*(1), 39–50. https://doi.org/10.5406/amerjpsyc.123.1.0039

Statistical Research and Applications Branch. (2011). *Joinpoint regression program (version 3.5.2) [computer software]*. National Cancer Institute.

Sundet, J., Barlaug, D., & Torjussen, T. (2004). The end of the Flynn effect? A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. *Intelligence, 32*(4), 349–362. https://doi.org/10.1016/S0160-2896(04)00052-2

Swing, E. L., Gentile, D. A., Anderson, C. A., & Walsh, D. A. (2010). Television and video game exposure and the development of attention problems. *Pediatrics, 126*(2), 214–221. https://doi.org/10.1542/peds.2009-1508

Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn effect in reverse. *Personality and Individual Differences, 39*(4), 837–843. https://doi.org/10.1016/j.paid.2005.01.029

Teasdale, T. W., & Owen, D. R. (2008). Secular declines in cognitive test scores: A reversal of the Flynn effect. *Intelligence, 36*(2), 121–126. https://doi.org/10.1016/j.intell.2007.01.007

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Twenge, J. M., & Campbell, W. K. (2001). Age and birth cohort differences in self-esteem: A cross-temporal meta-analysis. *Personality and Social Psychology Review, 5*(4), 321–344. https://doi.org/10.1207/S15327957PSPR0504_3

United Nations. (2014). *World economic situation and prospects 2014*. United Nations. https://www.un.org/en/development/desa/policy/wesp/wesp_current/2014wesp_country_classification.pdf.

Vedechkina, M., & Borgonovi, F. (2021). A review of evidence on the role of digital technology in shaping attention and cognitive control in children. *Frontiers in Psychology, 12*, 611155. https://doi.org/10.3389/fpsyg.2021.611155

Wieselmann, J. R., Roehrig, G. H., & Kim, J. N. (2020). Who succeeds in STEM? Elementary girls' attitudes and beliefs about self and STEM. *School Science and Mathematics, 120*(5), 297–308. https://doi.org/10.1111/ssm.12407

Wongupparaj, P., Wongupparaj, R., Kumari, V., & Morris, R. G. (2017). The Flynn effect for verbal and visuospatial short-term and working memory: A cross-temporal meta-analysis. *Intelligence, 64*, 71–80. https://doi.org/10.1016/j.intell.2017.07.006

Zheng, F., Gao, P., He, M., Li, M., Wang, C., Zeng, Q., Zhou, Z., Yu, Z., & Zhang, L. (2014). Association between mobile phone use and inattention in 7102 Chinese adolescents: A population-based cross-sectional study. *BMC Public Health, 14*(1), 1022. https://doi.org/10.1186/1471-2458-14-1022