universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Domain-specificity of Flynn effects in the CHC-model: Stratum II test score changes in Germanophone samples (1996-2018)“

verfasst von / submitted by

Alexandros Lazaridis, BSc.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2022 / Vienna 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 840
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Psychologie UG2002
Betreut von / Supervisor:	Ass.-Prof. PD Mag. Dr. Jakob Pietschnig, FHEA

Abstract

Generational IQ test score changes (the Flynn effect) were globally positive over large parts of the 20th century. However, accumulating evidence of recent studies show a rather inconsistent pattern of the Flynn effect in past decades. Patterns of recently observed test score changes appeared to be markedly different in strength and even signs between countries and domains. Because of between-study design differences and data availability in terms of differing IQ domains, so far it is unclear if these inconsistencies represent a consequence of differences in Flynn effect trajectories between countries, covered time-spans, or investigated IQ domains. Here, we present data from 36 largely population-representative Germanophone standardization samples from 12 well-established psychometric tests from 10 stratum II IQ domains from 1996 to 2018, thus providing a comprehensive assessments of domain-specific changes according to the Cattell-Horn-Carroll model of intelligence. Examination of both raw score and measurement-invariant latent mean changes yielded an erratic pattern of results, indicating positive (comprehension-knowledge, learning-efficiency, domain-specific knowledge), negative (working memory capacity), stagnating (processing speed, reading and writing), and ambiguous (fluid reasoning, reaction and decision speed, quantitative knowledge, visual processing) Flynn effects on stratum II domains. Our results show that the Flynn effect in Germanophones has become increasingly volatile over the past three decades, showing erratic patterns of test score changes between and within domains, even when limiting interpretations to measurement invariant and population representative assessments. The present findings may be indicative of an impending stagnation of the Flynn effect**.**

Introduction

The Flynn Effect

In 1984, James Flynn published results of his analyses of US standardization samples of the Stanford-Binet and Wechsler scales, two of the best-known intelligence tests, both widely-used in practice and research. His observation that average IQ scores of Americans had been increasing from 1932 to 1978 had an enormous scientific impact and has become known as the Flynn effect (Flynn, 1984). Subsequent studies have produced a large body of scientific literature pertaining to cross-temporal changes in average IQ test performance. While the strength of these changes varies substantially across different intelligence domains (fluid IQ gains have typically been found to be stronger than crystallized gains) and countries,

they were shown to have been predominantly positive in the twentieth century, but negatively associated with psychometric *g* (for an overview, see Pietschnig & Voracek, 2015). Over most of the 1900s, IQ increases averaged about 3 points per decade, although the strength of gains appeared to decline from the late 1980s to the early 2010s globally (Pietschnig & Voracek, 2015). Implications of this observation are relevant for both science and practice. Norm obsolescence as a result of these changes may have important consequences for the personal life or the professional career of test-takers in the course of cognitive ability assessments (e.g., Trahan et al. 2014).

A linear increase in average performance implies that a given generation would outperform any previous one. Also, the difference in performance between cohorts would necessarily become more pronounced the greater their distance in time was. However, recent findings indicate that (i) performance changes vary across intelligence domains and countries and (ii) their change trajectories are not linear (Pietschnig & Voracek, 2015). Moreover, evidence for directional changes of the Flynn effect have been reported in several studies, suggesting a deceleration (USA; Rindermann & Thompson, 2013) or stagnation of the Flynn effect in some countries (Australia; Cotton et al., 2005) and even a reversal in others (e.g., Denmark, Finland, France, Estonia, Netherlands, Norway, UK; for a review see Dutton et al., 2016).

Despite intensive research, to date only few available studies have provided comprehensive assessments of domain-specific changes according to current IQ models (e.g., the Cattell-Horn-Carroll-model model of intelligence; see McGrew & Schneider, 2018). Domain-specific Flynn effect research is typically focused on fluid and crystallized intelligence changes with most evidence pertaining to changes in the Wechsler and Stanford-Binet tests. Accounts of IQ changes in other domains are available (e.g., spatial ability: Pietschnig & Gittler, 2015; working memory: Wongupparaj et al., 2017) but are typically reported in a rather piecemeal fashion because they are based on data from different countries across different time periods and therefore reported separately. This problem is not rooted in suboptimal research approaches but a consequence of the typically retrospective nature of Flynn effect studies which does not allow any control of the researcher about the domains that can be investigated. However, this means that domain differences beyond the well-established variations between crystallized and fluid IQ cannot be disentangled from time-period- or country-specific effects. Consequently, there is a lack of empirical evidence that fully represents the Flynn effect on an integrative model of human intelligence.

The Cattell-Horn-Carroll Model of Human Intelligence

The CHC-Theory is a well-established integrative model of human cognitive abilities, formulated by McGrew (1997) and subsequently revised and extended (McGrew, 2009; Schneider & McGrew, 2012; Schneider & McGrew, 2018). It was developed as a synthesis of Cattell's Investment-Theory (Cattell, 1963), the Horn-Cattell Gf-Gc theory (Horn & Noll, 1997) respectively, and Carroll's Three-Stratum Theory of cognitive abilities (Carroll, 1993). Hence, this theory intregrates the two arguably most prominent and significant models of human intelligence. In that, it directly and indirectly incorporates findings of more than a century of psychometric research, and in particular, factor analytical research, that have been pivotal to our current understanding of the structure of human intelligence (e.g. Spearman's *g*-factor; Thurstone's Primary Mental Abilities; Vernon's hierarchical model). Here, the CHC-Theory provides a theoretically and empirically well-supported comprehensive structure of intelligence, most suitable as a framework for the cognitive measurements included. Table 1 shows all of the consistently identified broad abilities and their definitions, classified in terms of a conceptual-functional grouping as provided by McGrew (2009).

Table 1

Grouping and definitions of the broad abilities in stratum-II of the CHC-Theory

Conceptual-functional grouping	Broad ability	Definition
Fluid reasoning	Fluid reasoning (*Gf*)	The use of deliberate and controlled mental operations to solve novel problems that cannot be performed automatically
Memory	Short-term memory (*Gsm*)	The ability to apprehend and maintain awareness of a limited number of elements of information in the immediate situation
	Long-term storage and retrieval (*Glr*) [a]	The ability to store and consolidate new information in long-term memory and later fluently retrieve the stored information through association
General speed	Processing speed (*Gs*)	The ability to automatically and fluently perform relatively easy or over-learned elementary cognitive tasks, especially when high mental efficiency is required
	Reaction and decision speed (*Gt)*	The ability to make elementary decisions and/or responses (simple reaction time) or one of several elementary decisions and/or responses (complex reaction time) at the onset of simple stimuli
Sensory	Visual processing (*Gv*)	The ability to generate, store, retrieve, and transform visual images and sensations

	Auditory processing (*Ga*) [b]	Abilities that depend on sound as input and on the functioning of our hearing apparatus. A key characteristic is the extent an individual can cognitively control the perception of auditory information.
Acquired knowledge	Comprehension-knowledge (*Gc*) [c]	The knowledge of the culture that is incorporated by individuals through a process of acculturation. A person's breadth and depth of acquired knowledge of the language, information and concepts of a specific culture, and/or the application of this knowledge.
	Reading and writing (*Grw*) [c]	The breadth and depth of a person's acquired store of declarative and procedural reading and writing skills and knowledge
	General (domain-specific) knowledge (*Gkn*)	The breadth, depth and mastery of a person's acquired knowledge in specialized (demarcated) subject matter of discipline domains that typically do not represent the general universal experiences of individuals in a culture (Gc)
	Quantitative knowledge (*Gq*)	The breadth and depth of a person's acquired store of declarative and procedural quantitative or numerical knowledge

Note: [a] in the latest revision Schneider & McGrew (2018) separated this ability into Learning efficiency (*Gl*) and Retrieval fluency (*Gr*), we resort to the former definition to be consistent with previous research; [b] not accounted for by the data; [c] crystallized intelligence according to Carroll (1993)

Another problem in Flynn effect research pertains to the population representativity of the samples that have been assessed. Much data in this context originates from systematic large-scale assessments, such as admission or aptitude tests for educational and military institutions. For example, military draftees are disproportionately overrepresented in the literature on IQ changes. While these draftees are typically representative for 18 year old men, they are necessarily not representative for sex or age. Similarly, most of the available Flynn effect studies so far have been based on comparatively young samples, with over 90% of samples of all Flynn effect primary studies showing mean ages under 38 (Pietschnig & Voracek, 2015). Other frequently used approaches such as cross-sectional assessments of a single sample (e.g., Pietschnig, Voracek, & Formann, 2011) may underestimate the Flynn effect because of well-known mechanisms of restricted variances. Investigations of the Flynn effect in population-representative samples are therefore needed.

Another problem in interpreting IQ changes as genuine changes of cognitive abilities pertains to the variance of measurements between cohorts. Measurement invariance means the extent to which a test measures the same cognitive construct in the same way in different groups. Establishing measurement invariance aims to answer the fundamental question whether observed changes in IQ test scores are due to genuine effects of changed cognitive abilities or if they may be a consequence of measurement artefacts, such as changes in test-taking behavior or item drift (for an overview, see Wicherts et al., 2004). Therefore, test score changes can only meaningfully be attributed to changes in the latent cognitive construct when measurement invariance has been established.

The Flynn Effect in Germanophones

In Germanophones (i.e., representing individuals from Austria, Germany, and Switzerland), accounts of the Flynn effect were so far for the most part consistent with the global trend, yielding gains of about 2.4 IQ points per decade from 1962 to 2000 (Pietschnig & Voracek, 2015). Unlike in most other countries, IQ gains in Austria were similar in size across crystallized and fluid intelligence (Pietschnig, Voracek, & Formann, 2010). In recent years, observed change patterns have become more erratic, indicating virtual stagnation of the Flynn effect in Germanophone preschoolers (Pietschnig et al., 2021) or even a reversal in spatial ability (Pietschnig & Gittler, 2015). However, to date it is unclear if these inconsistencies are due to differing sample characteristics (e.g., children vs. adults) and investigated domains (e.g., spatial vs. other tasks) or if they represent a genuine change in the trajectory of IQ test score changes. Consequently, it seems necessary to systematically

examine the Flynn effect in different IQ domains in population-representative samples to allow disentangling study design-specific effects from IQ test score changes. This may contribute to a better understanding of the seemingly erratic Flynn effect trajectory in Germanophones in particular and in other countries in general.

To this end, we analyze data from 36 Germanophone standardization samples of 12 psychometric tests comprising 17 subtests, that were obtained from a well-established test publisher in Austria. These tests assess 10 different intelligence domains that represent stratum II domains of the Cattell-Horn-Carroll (CHC) model (Schneider & McGrew, 2018), one of the currently most widely-accepted integrative conceptualizations of human intelligence. Measures included tests for Fluid reasoning (Gf), Visual processing (Gv), Learning efficiency (Gl), Working memory capacity (Gwm), Reaction and decision speed (Gt), Processing speed (Gs), Comprehension-knowledge (Gc), Reading and writing (Grw), General (domain-specific) knowledge (Gkn), and Quantitative knowledge (Gq). To allow a meaningful interpretation of observed IQ changes, wherever possible, between-cohorts measurement invariance was established and latent-score changes were calculated by means of multi-group confirmatory factor analysis (MGCFA). The resulting findings allow for a comprehensive interpretation of cognitive performance changes in the Germanophone population between 1996 and 2018.

Proposed Causes of the Flynn Effect

Possible explanations and proposed causes of the Flynn effect are manifold and complex, including amongst others, genetic, biological and behavioral factors, as well as environmental and societal changes. Underlying is the fundamental question whether changes in IQ test scores are an indication of improved latent cognitive abilities or, for example, a result of a demographic transition or an artefact of changes in test sophistication changes or test-taking behavior. Because of the magnitude and relatively short time periods, genetic factors seem to be unlikely, notwithstanding heterosis as a result of immigration (Jensen, 1989; Mingroni, 2007). While there are many different opinions among intelligence researcher as to what causes the Flynn effect, there is some consensus that it is predominantly driven by environmental and societal changes, such as improved education and social multipliers. Other hypotheses are supported, inter alia, by studies investigating secular changes in infants and preschoolers, proposing improved nutrition as a probable influence (e.g.: Lynn, 2009). Of particular interest is the extent to which the Flynn effect can be attributed to *g*. Pietschnig and Voracek (2015) coined the term of *g-ness* to categorize

cognitive ability measures with respect to their level of *g* saturation. They accordingly assigned measures of *crystallized intelligence* to the highest g-ness level, measures of *full-scale intelligence* to the "medium" level, and those of *fluid intelligence* the "low" level. The observation that test scores gains are typically more pronounced for measures of fluid intelligence than for those of crystallized intelligence implies that these changes are relatively little influenced by *g*, suggesting environmental changes rather than genetic or biological factors as causes for the Flynn effect (e.g. Nijenhuis and van der Flier, 2013). With regard to a possible reversal and decline of the Flynn effect, potential influencing factors are considered to be largely the same and probably also rooted in environmental changes (Menie & Penaherrera-Aguirre, 2018).

Methods

Participants

For this study, we obtained standardization samples of psychological tests from the database of the SCHUHFRIED GmbH, a well-established developer and publisher of psychometric tests in Germanophone countries. We aimed at selecting tests that would allow us to most closely represent the stratum II factors of the Cattell-Horn-Carroll model (McGrew, 2009; Schneider & McGrew, 2018). Participants in this study were a total of 12,459 Austrian residents that participated from 1996 to 2018 in test standardizations.

Procedure

All participants were tested in the test and research center of the SCHUHFRIED GmbH, at partner Universities, or in proctored online settings, using a stratified sampling plan based on the demographic variables sex and age. Age was stratified by 14 levels in increments of five years from ages of fifteen to 80+ (excepting 6 samples which did not include participants over 80; see Table 1). The expected distribution for stratification was obtained from census data for Austria, Germany, and Switzerland (European Commission, 2022) and matched accordingly. All persons were tested under supervision of a trained test administrator, participated in the study voluntarily, and received about 9USD per hour for study participation. Following the data collection, chi-squared-tests were calculated to compare the demographic distribution of the standardization samples and corresponding census data. For some instruments, several samples had been combined in the course of re-standardization into a common norm for the use by the test publisher. We separated these

samples according to their original survey period and subsequently compared their demographic distribution with the matching census data to clearly determine their representativeness for the respective time period. Table 2 provides an overview of included samples and their demographic distribution in terms of age and education.

Table 2

Sample characteristics

Test		Standardization	Age		Education Level [a]	
			min - max	mean (*SD*)	median	modal
Adaptive Matrices Test (*AMT*)		2002 [b]	18 - 81	37.50 (15.50)	4	4
		2005 [b]	18 - 81	36.22 (14.12)	3	4
		2014	17 - 82	44.42 (16.91)	3	3
Basic Intelligence Functions (*IBF*)		2005 [b]	14 - 77	38.14 (13.45)	3	3
		2011	16 - 85	44.57 (17.86)	3	3
		2018	14 - 83	41.84 (16.30)	4	4
Block-Tapping Test forwards (*CORSI*)		2008	15 - 89	44.85 (17.44)	3	3
		2013	16 - 85	45.52 (16.84)	3	3
Cognitrone (*COG*; 4 test-forms)	1	2007	14 - 95	44.27 (17.82)	3	3
		2017	15 - 91	46.95 (16.07)	4	4
	2	2000	13 - 83	47.31 (16.62)	3	3
		2011	16 - 85	44.05 (16.71)	3	3
	3	2001	18 - 88	45.80 (17.36)	3	3
		2016	16 - 95	48.68 (17.60)	3	3
	4	1996 [c]	17 - 80	46.16 (17.01)	-	-
		2016	17 - 83	47.91 (16.10)	4	4
English Language Skills Test (*ELST*)		2009	13 - 64	39.11 (13.75)	3	3
		2017	16 - 80	46.12 (16.26)	4	4
Mathematics in Practice (*MIP*)		2002 [b]	18 - 71	38.34 (12.98)	4	4
		2011	16 - 84	44.05 (17.31)	3	3
Raven's Standard Progressive Matrices (*SPM*)		2001	16 - 82	45.57 (17.09)	3	3
		2010	14 - 81	43.62 (16.87)	3	3
		2017	16 - 87	49.91 (17.02)	4	4
Reaction Test (*RT*)		1999 [b]	16 - 89	40.90 (17.11)	3	3
		2007 [b]	16 - 76	39.65 (13.32)	3	3
		2016 [b]	17 - 83	49.57 (15.95)	4	3

Test	Standardization	Age		Education Level [a]	
		min - max	mean (*SD*)	median	modal
Reading Comprehension Test (*LEVE*)	2007	12 - 82	42.78 (16.59)	3	3
	2016	15 - 82	46.12 (17.58)	4	3
Spatial Orientation (*3D*)	2005	15 - 91	44.28 (17.79)	3	3
	2014	16 - 83	45.60 (16.84)	3	3
	2016	16 - 82	46.47 (17.20)	4	3
Visualization (*2D*)	2004	18 - 91	45.20 (17.93)	3	3
	2010	16 - 81	43.15 (16.81)	3	3
	2013	17 - 83	48.67 (16.30)	4	3
Work Performance Series (*ALS*)	2003	17 - 60	37.75 (12.82)	3	3
	2008	15 - 75	39.14 (14.03)	3	3

Note: [a] highest completed education; levels classified according to EU-Standards: EU educational level 1 = no school-leaving qualification, EU educational level 2 = compulsory schooling or an intermediate secondary school but without completing vocational training, EU educational level 3 = completion of vocational training or a course at a technical college, EU educational level 4 = school-leaving qualification at university entrance level, EU educational level 5 = university degree. [b] Indicates samples that are not population representative in terms of age and sex. [c] no education level data was available for this sample.

Materials

We included standardization data from 17 (sub-)tests: three and four subtests of two intelligence test batteries, respectively, as well as 10 specific ability tests. For each subtest, multiple standardization samples (*k* range = 2-3, *Mdn* = 2) from different years were available. Initially we located a total of 59 samples, 23 of which were subsequently excluded. This was necessary either because of non-equivalence between tests, due to different test difficulty across forms, or because of inappropriate between-cohort intervals (i.e., time interval between samples < 2 years). Consequently, we included 36 standardization samples that had been recruited from 1996 to 2018 in our analyses, with an average timespan of 11 years that had elapsed between the initial and re-standardization of tests (*SD* = 4.66). Table 3 provides an overview over stratum II abilities, timeframes, included tests and samples, as well as their characteristics.

Table 3

Overview of included tests and samples

Test		Subtest/Construct [a]	Stratum II domain	Samples				Time-span
Adaptive Matrices Test (*AMT*)		Logical reasoning	Fluid reasoning (Gf)	Collection year	2002 [b]	2005 [b]	2014	12
				N	209	434	267	
Basic Intelligence Functions (*IBF*)		Numerical Intelligence (NIF-A) Numerical Intelligence (NIF-B) Visualization (VIZ) Long-term memory (LTM) Verbal Intelligence (VIF)	Fluid reasoning (Gf) Quantitative knowledge (Gq) Visual processing (Gv) Learning efficiency (Gl) Comprehension-knowledge (Gc)	Collection year	2005 [b]	2011	2018	13
				N	630	476	286	
Block-Tapping Test forwards (*CORSI*)		Short term storage	Working memory capacity (Gwm)	Collection year	2008	2013		5
				N	321	567		
Cognitrone (*COG*; 4 test-forms)	1	Concentration	Reaction and decision speed (Gt)	Collection year	2007	2017		10
				N	554	300		
	2			Collection year	2000	2011		11
				N	220	311		
	3			Collection year	2001	2016		15
				N	179	354		

Test	Subtest/Construct [a]	Stratum II domain	Samples				Time-span
4			Collection year	1996	2016		20
			N	492	318		
English Language Skills Test (*ELST*)	Reading comprehension (RC) Grammar (GR) Vocabulary (VO)	General (domain-specific) knowledge (Gkn)	Collection year	2009	2017		8
			N	274	304		
Mathematics in Practice (*MIP*)	Calculation capacity	Quantitative knowledge (Gq)	Collection year	2002 [b]	2011		9
			N	260	298		
Raven's Standard Progressive Matrices (*SPM*)	Logical reasoning	Fluid reasoning (Gf)	Collection year	2001	2010	2017	16
			N	244	319	340	
Reaction Test (*RT*)	Decision speed	Reaction and decision speed (Gt)	Collection year	1999 [b]	2007 [b]	2016 [b]	17
			N	550	246	363	
Reading Comprehension Test (*LEVE*)	Reading comprehension	Reading and writing (Grw)	Collection year	2007	2016		9
			N	362	344		
Spatial Orientation (*3D*)	Spatial orientation	Visual processing (Gv)	Collection year	2005	2014	2016	11
			N	356	361	357	
Visualization (*2D*)	Visualization	Visual processing (Gv)	Collection year	2004	2010	2013	9
			N	254	331	341	

Test	Subtest/Construct [a]	Stratum II domain	Samples			Time-span
Work Performance Series (*ALS*)	General performance	Processing speed (Gs)	Collection year	2003	2008	5
			N	302	303	

Note: [a] Test/construct label according to test-battery or manual; [b] sample is not population representative in terms of age and sex.

Adaptive Matrices Test (AMT; Hornke, Etzel, & Rettig, 2020)

The AMT is a non-verbal assessment of inductive reasoning. The theoretical basis for the construction of the test is provided by the CHC-Theory, which specifies *inductive reasoning* as "the ability to observe a phenomenon and discover the underlying principles or rules that determine its behavior" (Schneider & McGrew, 2012, pp. 112). The AMT consists of Raven's-typed figural matrices items, each of which consists of nine elements in a 3*3 matrix design, displaying eight geometrical forms with a particular relationship, and one missing element. The task for the respondent is to identify the relationships between these elements and select the correct ninth element from a set of eight alternatives. Integrated in this study are three test forms, all of which are administered in an adaptive form and vary in length and their measurement precision (standard error of measurement [reliability]: test forms 1 to 3 = 0.63 [0.70], 0.44 [0.83], and 0.39 [0.86]). The outcome variable for all of these forms is the estimated person parameter as an estimate of the subjects' ability for induction and the stratum II factor fluid reasoning (Gf) with respect to the CHC-Framework.

Basic Intelligence Functions (IBF; Blum et al., 2018)

The IBF is an intelligence test-battery which is based on a modified version of the intelligence theory of Thurstone and Thurstone (1941) and consists of four subtests, corresponding to the primary factors numerical intelligence, verbal comprehension, figural-spatial ability, and memory (Thurstone, 1938; Thurstone & Thurstone, 1941). The outcome for all subtests is operationalized as the sum of correct answers, and available as raw score and *z*-standardized variable of the raw score. Raw sum scores, are further differentiated within two subtests of the IBF due to multiple item formats. The numerical intelligence functions (NIF) and the verbal intelligence functions (VIF) of the IBF, each consist of different tasks by means of two different item formats. The raw sum scores for both subtests are calculated accordingly for each item format separately. Here, we will include the raw sum scores in our analyses, to differentiate between different tasks in the results. In addition to the sum scores for each subtest, a combined overall score is computed, based on the weighted factor scores of all four subtests. This global score can be interpreted as an indication of the respondents' general cognitive ability (*g*). Presently, we included data from the standard version of the test.

The subtest numerical intelligence functions (NIF) of the IBF consists of two parts with different types of item formats and includes a total of 40 items in the standard test form with an average working time of about 25 minutes. The subtest measures inductive reasoning by means of simple number sequence tasks, and calculation capacity by means of basic

arithmetic tasks in text format. In the first part (NIF-A), respondents need to identify a logical rule in a sequence of seven numbers and need to specify the correct subsequent number. In the second part (NIF-B), respondents are presented with practical problems that are embedded in a text. These problems involve basic arithmetic operations and require the respondent to identify relationships between them to solve the problem. Integrated in the CHC-taxonomy, the outcome of the first part (NIF-A) can be understood as an indication of the primary factor induction, and thus, the stratum II factor fluid reasoning (Gf). The outcome of the second part (NIF-B) is an indication of the primary factor mathematics achievement and thus, the stratum II factor quantitative knowledge (Gq).

The subtest visualization (VIZ) of the IBF is a non-verbal assessment designed to measure abilities associated with visualization or mental rotation. The subtest consists of 17 items with an overall working time of about 18 minutes in the standard test form. Respondents are presented with a reference figure that is a three-dimensional cube, and a set of alternative figures that are another six three-dimensional cubes, and asked to decide which one of these alternatives, if any at all, is a rotated version of the reference cube. The underlying cognitive task is to form a corresponding mental representation of the reference figure, mentally rotate it and compare it with the alternatives (Just & Carpenter, 1985). Embedded in the taxonomy of the CHC-theory we can understand the associated ability as a measure of the primary factor visualization and the stratum II factor visual processing (Gv).

The subtest long-term memory (LTM) of the IBF is divided into two phases. In the memorization phase, items containing specific information are presented and need to be memorized in seven minutes. After 20 minutes, subjects need to answer questions about these items in a subsequent reproduction phase, for which they have a time limit of six minutes. The standard form of this subtest consists of 20 items and participants take 13 minutes of total working time. Outcomes from this subtest can be viewed as measurements of the primary factor meaningful memory, speed of lexical access, and naming facility & word fluency and therefore of the stratum II factor learning efficiency (Gl).

The subtest verbal intelligence functions (VIF) of the IBF is an assessment of verbal comprehension, defined as the ability to understand the meaning and implications of words and to apply them appropriately in a conversational context (Thurstone, 1938; Thurstone & Thurstone, 1941). This subtest consists of 35 items with a total working time of about 11 minutes and two different types of item formats. In one part of the subtest (VIF-A), respondents need to complete a sentence by selecting a correct word or phrase out of a set of alternatives. In this task, respondents must apply general knowledge in a way that requires an

understanding of the meaning of words as well as their implications in a particular context. In a second part (VIF-B), respondents are confronted with verbal analogies. For each item, three words are presented. The first and second word are presented as a pair that has a certain relationship. The participants then need to establish an analogous relationship between a third word with a fourth one that has to be selected from a set of five alternatives. Both tasks involve abilities associated with the primary factors language development, lexical knowledge, and general verbal information. Outcomes of this subtest can therefore be understood as a measurement of the stratum II factor comprehension-knowledge (Gc).

Block-Tapping Test (CORSI; Schellig, 1995)

The CORSI is the computerized version of the original Block-Tapping Task developed by Corsi (1972). This instrument provides a non-verbal assessment of visuospatial working memory and is widely-used, both in research and practice. The Block-Tapping Test is a non-verbal analogue to the paradigm of the digit span (Hebb, 1961) as realized for instance in the Wechsler scales (e.g. WAIS-R; Wechsler, 1981) and is considered a gold standard for the assessment of spatial components of the working memory (Baddeley, 2001; 2003). We included the standard form of the CORSI, which contains a total of 28 items that represent sequences that become longer as difficulty increases. Working time ranges from 10 to 15 minutes. Respondents are presented with nine irregularly positioned blocks that light up in a particular sequence. This sequence then needs to be repeated by the respondent. The available outcome for our norm samples was immediate block span (UBS) forwards, operationalized as the longest sequence that was correctly reproduced at least twice. This measures the short-term storage capacity for a sequence of spatial details and can accordingly be assigned to the primary factor visuo-spatial short-term storage and the stratum II factor working memory capacity (Gwm).

Cognitrone (COG; Schuhfried, 2020a)

The COG is a non-verbal assessment of abilities associated with concentration. The construct of concentration can be understood as the perception and processing of information for different tasks (Schmidt-Atzert, Büttner, & Bühner, 2004) and more specifically, the intentional and controlled coordination of cognitive activities (Westhoff & Hagemeister, 2005). The instrument includes eight different test forms which differ in the concrete administration conditions with regard to item working times, overall working time and number of items. Included in this study are the two parallel forms (S1, S2), with flexible

working time and 200 items each, a longer test form that is administered with flexible working time with 308 items (S3), and another one, with restricted working time for each of its 200 items (S4; 1.8 seconds time limit per item). The design of all these forms includes multiple simultaneously presented stimuli which must be visually compared. Respondents are required to compare a specific geometric figure with four alternatives and decide then whether these include any congruent or only incongruent alternatives (i.e., representing a mental comparison paradigm; Westhoff & Hagemeister, 2005). The outcome variable for the first three forms with flexible working time is the mean time of correct rejections, defined as the average time a respondent needs to decide that a comparison figure is not identical to any of the alternative figures. For the test form with limited working time, the outcome is operationalized as the sum of correct reactions and the sum of incorrect reactions. With regard to the taxonomy of the CHC-theory we can attribute the outcome of the COG to the primary factor mental comparison speed and the stratum II factor reaction and decision speed (Gt).

English Language Skills Test (ELST; Janous, Ortner, & Lick, 2020)

The ELST is a multi-dimensional, adaptive assessment of English as a foreign language comprising the subtests reading comprehension (RC), grammatical sensitivity (GR), and vocabulary (VO). Here, we integrated data from the adaptive form of the test which takes overall about 40 minutes of working time. In the subtest reading comprehension (RC) of the ELST, participants have to read a text and subsequently need to indicate if statements about this text are accurate or not. In the subtest grammatical sensitivity (GR) of the ELST, respondents are shown incomplete sentences and asked to select one or more words from a set of alternatives to complete the sentence in a grammatically correct manner. In the subtest vocabulary (VO) of the ELST, respondents are presented with simple sentences and asked to select the correct synonym or antonym for a specific word from a set of four alternatives. The outcome variable for all three subtests is the estimated person parameter as an indicator of the subjects' foreign-language proficiency, and therefore a measurement of the stratum II factor general domain specific knowledge (Gkn).

Mathematics in Practice (MIP; Bratfisch, Hagman, & Egle, 2018a)

The MIP is an assessment of basic arithmetic operations in the context of real-life situations and designed as a measure of calculation capacity with regard to Thurstone's Primary Mental Abilities (1938). This ability is distinct from others such as numerical ability and mathematical ability in that it concerns practical aspects of real-life mathematical

problems rather than the general understanding of underlying symbol or rule systems. Respondents are presented with "daily-life" calculation tasks including logical, numerical and verbal components, that require basic arithmetic operations (addition, subtraction, multiplication and division). Here, we included data from one test form of the MIP that contains 20 items without time limit. The outcome of this test, operationalized as the number of fully solved items, represents a measurement of the primary factor mathematics achievement, which is part of the stratum II factor quantitative knowledge (Gq).

Raven's Standard Progressive Matrices (SPM; Raven, Raven, & Court, 2018)

The SPM is a non-verbal assessment of cognitive abilities associated with abstract reasoning. The original version of the SPM (Raven & Court, 1938) was published in 1938 together with a form including colored matrices (CPM) and advanced matrices (APM). While only few changes were since then made to the design, the test has been extended with a series of parallel and special forms for various application purposes, and with a form including several difficult items in response to the Flynn effect and the accompanying eroding differentiability of the test in the upper performance range (SPM+; Raven, Styles & Raven, 1998). These tests have since been, and still are, among the most frequently used psychometric instruments in both, research and practice (Van de Vijver, 1997; Evers, 2011). Here, we included data from the standard form of the SPM, which consists of 60 figural matrices items, each of which consists of nine elements in a 3*3 matrix, displaying eight geometrical forms in a particular pattern, and one missing element. The task for the respondent is to identify the pattern and select the correct ninth element from a set of six or eight alternatives. The test takes about 47 minutes of total working time and is suitable for application with participants starting from five years. According to Raven, Raven, and Court (2018), the instrument allows for a measurement of eductive ability, defined as the ability to extract and understand information from a complex situation. Embedded in the CHC-taxonomy, the outcome, operationalized as the raw sum of correct answers, is an indicator of the primary factor induction and the stratum II factor fluid reasoning (Gf).

Reaction Test (RT; Schuhfried, 2020b)

The RT is a non-verbal assessment of reaction speed, and more precisely, the ability to react under simple stimulus conditions. This ability can be defined as the speed with which very easy decisions or judgements are made in the context of sequentially presented stimuli (Schneider & McGrew, 2018). Schneider and McGrew (2012; 2018) propose on the level of

primary factors further differentiation and specify amongst others the narrow abilities *simple reaction time* and *choice reaction time*. While simple reactions require only reaction to a single stimulus, choice reactions require a selection between multiple stimuli and reaction alternatives. The RT includes eight test forms which differ from each other by type and presentation mode of the stimuli. We included here data from one test form, which contains 48 items for a total working time of about six minutes. In this test form, a sequence of different visual stimuli, an acoustic stimulus, as well as combinations of both are presented. Respondents are required to react to one specific combination of acoustic and visual stimuli only (choice-reaction). Upon detection of a relevant stimulus, reaction is required in a twofold way, akin to the Jensen-box: first, by letting go of a touch-sensitive button on which a finger is resting on, and second, by pressing a separate button. This allows for a measurement of both, a cognitive and a motor component of the ability, accounted for by the outcome variables reaction speed and motor speed. Both variables are Box-Cox transformed mean times, to ensure optimal representation of the central tendency. Measurements of this test can be assigned to the primary factor choice reaction time and the stratum II factor reaction and decision speed (Gt) of the CHC-Theory. This time-critical test is predominantly used in the context of traffic psychology and other safety assessments, as well as in the area of sport psychology.

Reading Comprehension Test (LEVE; Proyer, Wagner-Menghin, & Grafinger, 2018)

The LEVE is an assessment of reading comprehension, including the components word comprehension, vocabulary, sentence comprehension, text comprehension, and memory. We included a 16-item form with an overall working time of approximately 22 minutes. In this form, respondents are required to read a multi-paged text without time constraints but without the possibility to page back. Subsequently, the respondent needs to answer multiple choice questions about the text. An index for overall ability of reading comprehension is subsequently calculated. The LEVE measures the primary factor reading comprehension and belongs therefore to the stratum II factor reading and writing (Grw).

Spatial Orientation (3D; Bratfisch, Hagman, & Egle, 2018b)

The 3D is a non-verbal assessment designed to measure spatial ability as defined by Thurstone (1938) and more specifically, spatial relations and spatial orientation, as the ability to perceive the environment from a new position and from a different perspective (Lohman, 1989). Here, we included data from the standard test form of the 3D which is restricted by a

fixed overall duration of three minutes and includes a total of 30 items. In each of these items, the respondent is presented with a three-dimensional structure compiled of several building blocks on the one hand, and a set of four alternative structures on the other hand, based on blocks of the same shape and size. The task for the respondent is to determine which of these alternatives represents the view of the reference figure from a different perspective as indicated by an arrow. Items and instructions of the 3D are typical for the assessment of mental rotation. While there is some evidence for a primary factor spatial relation (for an overview see: Hegarty & Waller, 2004), separate from visualization, it is debatable whether abilities associated with the measurement provided by the operationalization and instructions of the 3D are distinct from the latter one. Any assessment of spatial relations implies a measurement of such skills that require a person to imagine different perspectives or orientations of themselves in relation to an object. The distinction is the underlying spatial transformation, where visualization relies on object-based transformations (e.g. mental rotation) and spatial relation is based on egocentric spatial transformation (e.g. perspective taking) (Thurstone, 1950). Lohman (1979) claimed that markers of spatial relation such as the widely-used the Guilford-Zimmerman Spatial Orientation Test (Guilford & Zimmerman, 1948) do not measure a factor distinct from visualization and also Carroll (1993), in his seminal work, could not find any evidence for a separability of these two factors. The items and instructions of the 3D, while designed to measure spatial relations, resemble closely those of markers for mental rotation and are therefore unlikely to evoke any solution strategies distinct from such tests. For the purpose of this study, we therefore locate this measurement accordingly within the framework of the CHC-theory as a measurement of visualization and the secondary factor visual processing (Gv).

Visualization test (2D; Bratfisch, Hagman, & Egle, 2018c)

The 2D is a non-verbal assessment of spatial ability. Respondents of the 2D are shown a two-dimensional horizontal bar, that is missing a part. The respondents need to select a segment from a number of alternatives to complete the bar. The 2D consists of 22 items and takes approximately six minutes to complete. The task requires the participant to mentally transform a two-dimensional object (Arendasy & Sommer, 2010) or the ability to perceive and mentally transform complex patterns (Schneider & McGrew, 2018). These abilities are associated with the primary factor visualization and the stratum II factor visual processing (Gv).

Work Performance Series (ALS; Schuhfried, 2018)

The ALS is the computerized version of a continuous work test, integrating the principle of continuous addition as prototyped by Pauli (1921). The respondent's task is to perform as many additions of two simultaneously presented numbers as possible. Here, we included data from a test form which comprises 20 partial intervals, which take one minute each. The outcome of this test is operationalized by means of two variables. The number of items answered represents the total number of calculations performed, thus indicating performance speed, while error percentage, calculated as the number of incorrect solutions as a percentage of the total number of items answered, is indicative of performance quality. Embedded in the CHC-Framework this measurement can be assigned to the primary factor number facility and to the stratum II factor processing speed (Gs).

Statistical Analysis

Preliminary analyses were conducted to assess the correctness, completeness, and representativeness of the raw data. For this purpose, demographic distributions of all samples were compared with the documentation provided in the corresponding test manuals. When test manuals did not provide the necessary information, age- and sex-specific distributions of the raw data were compared with the census data for the corresponding Germanophone population by means of chi-squared-tests. All standardization samples were included in subsequent analyses.

In our primary analyses, we focused on cross-temporal test performance changes across cohorts. For each (sub-)test, this refers to changes in average performance between standardization samples, collected at different time points. For this purpose, we calculated pairwise standardized mean differences (Cohen *d*) between the raw scores or ability parameters across all samples of a given subtest. To facilitate the interpretation of resulting changes, we recoded all values so that larger means are indicative of better performance. This mean that in our results, positive *d*-values always represent performance increases (i.e., positive Flynn effects) and negative *d*s performance decreases over time (i.e., negative Flynn effects).

Resulting standardized mean differences were indicative of the strength of changes over time. We interpret effect size changes according to the well-established thresholds by Cohen (1988; absolute *d*s = 0.2, 0.5, and 0.8 representing the lower threshold for small, moderate, and large effects, respectively). *T*-tests, ANOVA's, and Bonferroni-corrected post-hoc tests were performed to establish formal significance of changes across all pairwise group

comparisons. Furthermore, to investigate potential sex effects on the Flynn effect, a series of ANCOVAs was calculated by entering sex as a covariate whilst controlling for age and education and examining interactions between sex and cohort.

To corroborate the results from the pairwise comparisons of raw scores, we assessed measurement invariance of tests across cohorts. Measurement invariance pertains to whether the results of the operationalization of a construct have the same meaning under different conditions (Meade & Lautenschlager, 2004), and thus provides support as to whether changes in an outcome can be meaningfully interpreted across groups (Putnick & Bornstein, 2016). We established stability of measurements over time, by assessing if a set of items measured the same construct with identical precision across different cohorts (Kline, 2015) by means of multi-group confirmatory factor analyses (MGCFA).

The implementation of MGCFA imposes certain requirements, such as availability of item-level data and identical items across conditions. Because of that, we were unable to perform those analyses for some of the included tests, because item-level data were either unavailable, contained different items due to adaptive administration, and differing (albeit equivalent) test forms. Therefore, data from the AMT, CORSI, ELST, as well as samples 2005 and 2013 from the IBF had to be excluded from measurement invariance analyses. Data from ALS, RT, and test forms S1, S2, and S3 of COG were analyzed following the classical approach for continuous indicators, by means of imposing increasingly restrictive parameter constraints across groups.

We progressively established levels of measurement invariance, starting with configural invariance, then weak invariance, strong invariance, and strict invariance (Wu, Li, & Zumbo, 2007). In this procedure, the configural (i.e., baseline) model, is defined by restricting only the configuration of the latent construct to be equal across groups. This indicates whether an equal number of factors and a correspondence between factors and indicators in all groups can be established. All other parameters are freely estimated within each group. Subsequently, these parameters are stepwise restricted in different models, resulting in different levels of measurement invariance.

This procedure is not appropriate for ordered categorical indicators, due to the absence of inherent scales of observed values. It has been argued that any identification condition of the configural model defines the scale of each latent continuous response, providing each with a mean and a variance for further analyses (Wu & Estabrook, 2016). These conditions become binding for the estimates of the measurement parameters and certain restrictions can therefore

not be tested in isolation. The procedure for testing measurement invariance for categorical indicators therefore differs from continuous ones.

For dichotomous data, a model with invariant thresholds and loadings is statistically equivalent to a baseline model (Wu & Estabrook, 2016). Consequently, this model served as identification condition of the baseline for data from the 2D, 3D, COG-S4, IBF, LEVE, and SPM. Subsequently, strict invariance was established by further imposing group-equality constraints on residual variances. For continuous data we defined strong invariance as the minimum requirement to ensure meaningful interpretation of group-wise changes in means, while for dichotomous items, this requirement necessarily had to be defined at the level of strict invariance. For both, continuous and dichotomous data, model comparisons were evaluated by inspecting CFI changes between models. If changes exceed 0.01 between two models, it indicates that measurement invariance cannot be established at the level of the more restrictive model (Cheung & Rensvold, 2002). In cases when full measurement invariance could not be accounted for, we planned to establish partial invariance by iteratively freeing indicators until changes in CFI dropped below 0.01. Based on the resulting models, we estimated latent factor scores and means, to investigate changes in the latent construct. We provide IQ changes per decade (*DIQ*) for raw as well as latent mean score-based changes. All statistical analyses were conducted in the Open Source Software environment R (R Core Team, 2019). Measurement invariance analyses were conducted by means of the package lavaan (Rosseel, 2012).

Results

Preliminary analyses revealed that out of the 36 included standardization samples, 29 were representative of the German-speaking population of their respective period in terms of age and sex. In the course of these analyses, a total of eight samples were examined, seven of which were not representative in terms of age and sex (see Table 4). All standardization samples were included in subsequent analyses.

Table 4

Results of chi-squared-tests for age- and sex-specific distribution

Test	Sample	x^2 (df, N)	*p*
AMT	2002	174.23 (27, 209)	<0.001
AMT	2005	238.16 (27, 434)	<0.001
AMT	2014	30.96 (27, 321)	0.273
IBF	2005	205.34 (27, 630)	<0.001
MIP	2002	137.52 (27, 260)	<0.001
RT	1999	155.46 (27, 550)	<0.001
RT	2007	90.33 (27, 246)	<0.001
RT	2016	61.05 (27, 363)	<0.001

Our primary analyses revealed performance change trajectories that were differentiated according to cognitive domains and tests. Pairwise comparisons across standardization samples resulted in standardized test score changes ranging from small and non-significant to large and significant effects (*p* range = <.001 to .98; absolute *d* range = <0.01 to 0.96). Raw and latent mean score-based cross-temporal changes were broadly consistent in strength and direction. We successfully established strong (continuous values) and strict (dichotomous responses) measurement invariance at the predefined minimum level for all comparisons that were eligible for MGCFA (excepting COG-S4 for which only configural but not strict invariance could be established due to model non-convergence; summary fit statistics for all measurement invariance calculations are provided in appendix C). Therefore, we focus on reporting changes in latent scores unless indicated otherwise. Table 5 provides results of the pairwise cohort comparisons across all (sub-)tests. All our reported Cohen *d*-values with positive signs represent IQ score increases whilst those with negative signs represent decreases.

Table 5

Test score changes of pairwise cohort comparisons for raw and latent score changes reported in Cohen's d and DIQ

Domain	Outcome	Interval	Raw scores		Latent scores	
			d	*DIQ*	*d*	*DIQ*
Adaptive Matrices Test (AMT)						
Fluid reasoning (Gf)	sum score	2002 [a] -2005 [a]	-0.05	-2.31	-	-
		2005 [a]-2014	0.02	0.41	-	-
Basis Intelligence Functions (IBF)						
Full-scale IQ (*g*)	Weighted average (4 subtests)	2005 [a]-2011	0.26**	6.44	-	-
		2011-2018	0.67**	14.42	0.69**	14.75
Fluid reasoning (Gf)	sum score (NIF-A)	2005 [a]-2011	0.35**	8.70	-	-
		2011 - 2018	0.59**	12.72	0.60**	12.75
Quantitative knowledge (Gq)	sum score (NIF-B)	2005 [a]-2011	0.24**	6.04	-	-
		2011 - 2018	0.50**	10.67	0.50**	10.78
Visual processing (Gv)	sum score (VIZ)	2005 [a]-2011	-0.07	-1.78	-	-
		2011-2018	0.74**	15.79	0.74**	15.84
Learning efficiency (Gl)	sum score (LTM)	2005 [a]-2011	0.31**	7.77	-	-
		2011-2013	0.40**	8.55	0.40**	8.55
Comprehension-knowledge (Gc)	sum score (VIF-A)	2005 [a]-2011	0.23**	5.63	-	-
		2011-2018	0.47**	10.16	0.49**	10.49
	sum score (VIF-B)	2005 [a]-2011	0.26**	6.51	-	-
		2011-2018	0.58**	12.43	0.59**	12.69
Block-Tapping Test (CORSI)						
Working memory capacity (Gwm)	longest sequence correctly reproduced	2008-2013	-0.48**	-14.44	-	-
Cognitrone (COG; test forms: S1, S2, S3, S4)						
Reaction and decision speed (Gt)	reaction time (S1)	2007-2017	0.10	1.45	0.11	1.65
	reaction time (S2)	2000-2011	0.003	0.04	0.03	0.35
	reaction time (S3)	2001-2016	0.35**	3.47	0.35**	3.45

	sum score - correct (S4)	1996-2016	0.15*	1.12	-	-
	sum score - wrong (S4)	1996-2016	0.96**	7.17	-	-
English Language Skill Test (ELST)						
General (domain-specific) knowledge (Gkn)	person parameter (Reading compr.)	2009-2017	0.43**	8.07	-	-
	person parameter (Grammar)	2009-2017	0.61**	11.41	-	-
	person parameter (Vocabulary)	2009-2017	0.64**	12.02	-	-
Mathematics in Practice (MIP)						
Quantitative knowledge (Gq)	sum score	2002 [a]-2011	-0.33**	-5.50	-0.33**	-5.42
Raven's Standard Progressive Matrices (SPM)						
Fluid reasoning (Gf)	sum score	2001-2010	-0.002	-0.03	0.003	0.06
		2010-2017	0.33**	6.99	0.32**	6.77
Reaction Test (RT)						
Reaction and decision speed (Gt)	reaction time	1999 [a]-2007 [a]	-0.08	-1.45	-0.07	-1.36
		2007 [a]-2016 [a]	-0.38**	-6.37	-0.37**	-6.13
	motor time	1999 [a]-2007 [a]	-0.25*	-4.63	-0.26*	-4.84
		2007 [a]-2016 [a]	-0.15	-2.43	-0.15	-2.54
Reading Comprehension Test (LEVE)						
Reading and writing (Grw)	sum score	2007-2016	0.01	0.18	0.01	0.16
Spatial Orientation (3D)						
Visual processing (Gv)	sum score	2005-2014	-0.35**	-5.87	-0.39**	-6.55
		2014-2016	0.26**	19.64	0.29**	21.6
Visualization (2D)						
Visual processing (Gv)	sum score	2004-2010	-0.02	-0.45	-0.07	-1.64
		2010-2016	0.11	2.84	0.14	3.38
Work Performance Series (ALS)						
Processing speed (Gs)	items answered	2003-2008	-0.07	-2.18	-0.07	-2.18
	error percentage	2003-2008	-0.16*	-4.89	-0.16*	-4.83

Note: changes in IQ per decade (DIQ) can be calculated with the following formula: DIQ (interval) = ((*d* * 15) / interval); [a] Sample is not population representative for age and sex. * = $p < .05$; ** = $p < .001$

For the AMT, we observed non-significant and mostly negligible changes in test scores, with different signs in the two observed intervals. Between standardization samples of 2002 and 2005, changes in average performance were non-significant, negative, and trivial (*d* = -0.05; *p* = .59), while between 2005 and 2014 changes were positive, non-significant, and trivial (*d* = 0.02; *p* = .75), thus indicating virtually no change over the entire observed period from 2002 to 2014 (*d* = -0.02; *p* = .82; Figure 1, panel A).

For the IBF, we observed almost exclusively performance increases between 2005 and 2018, with trivial-to-moderate effects across different cohorts (*d*s ranging from -0.07 to 0.74) and a tendency towards larger gains in later years. This pattern was largely consistent over the observed time-span and across all four subtests (numerical intelligence, visualization, verbal intelligence, and long-term memory) as well as the global score (i.e., the weighted average of all subtests) as a measure of full-scale cognitive ability.

However, examination of incremental changes between restandardization samples showed, that the changes appeared to be non-linear. In the time interval between 2005 and 2011, we observed significant small increases in test performance for full-scale cognitive ability (g), fluid reasoning (Gf), quantitative knowledge (Gq), learning efficiency (Gl), and comprehension-knowledge (Gc; *d*s ranging from 0.21 to 0.30). The only exception to this was visual processing (Gv) where we observed a trivial and non-significant test performance decrease (*d* = -0.07; *p* = .76). Results from the comparisons of samples in the subsequent interval from 2011 and 2018 revealed substantial increases across all domains, with small-to-moderate effects (*d*s ranging from 0.40 to 0.74). Over the entire observed time period from 2005 to 2018, changes based on raw scores were consistently significant and positive, with moderate-to-large effects across domains (*d*s = 0.67 to 0.91; *p*s <.001). Changes based on latent means were only available for the pairwise comparisons of cohorts in 2011 and 2018

but showed virtually identical outcomes with significant increases in test performance across all domains (*d*s ranging from 0.40 to 0.74; Figure 1, panels B to H).

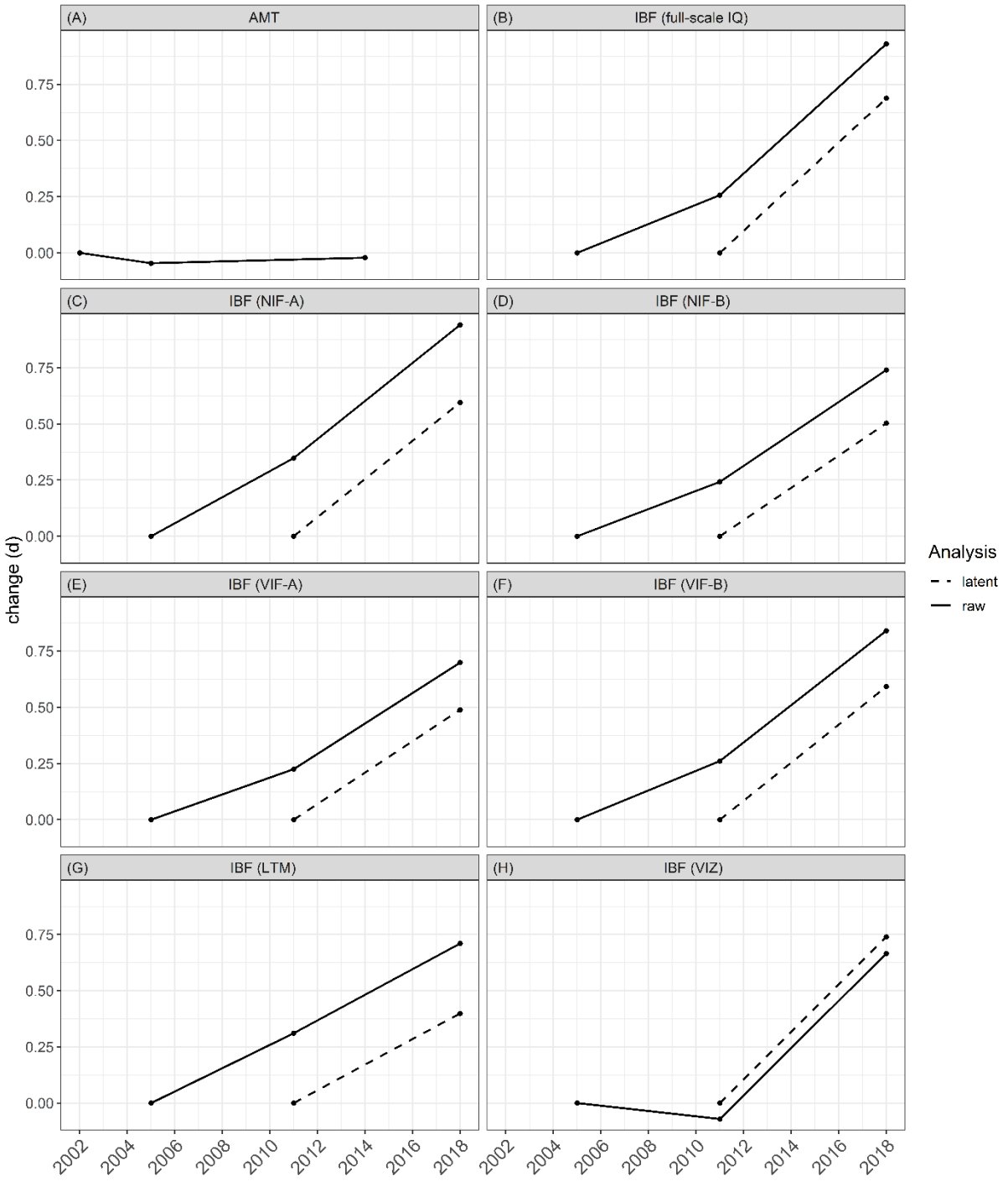

Figure 1: Test score change trajectories for the AMT (A) and IBF (B - H).

For the CORSI, a comparison of standardization samples from 2008 and 2013 revealed a significant and small-to-moderate decrease in test performance (d = -0.48; p < .001; Figure 2, panel A).

For the COG, comparisons of standardization samples yielded performance increases, with trivial-to-large effects (*d*s = 0.03 to 0.96) from 1996 to 2016. These changes, while consistent in sign, differed considerably in strength across test forms of the COG. Two test forms with flexible working time (S1, S2) revealed non-significant and mostly trivial increases in average performance for the main outcome mean reaction time between 2007 and 2017 (d = 0.11; p = .13), as well as 2000 and 2011 (d = 0.03; p = .77), respectively. However, the third test form without working time limit (S3), showed a small increase between 2001 and 2016 (d = 0.35; p < .001). Raw scores of the fourth test form, which was administered with a restricted working time (S4), increased significantly between 1996 and 2016, for both sum of correct as well as incorrect answers. Effect sizes of both outcomes varied considerably and were substantially larger for the sum of incorrect answers (d = 0.96; p < 001), than for the sum of correct answers (d = 0.15; p = .04; Figure 2, panels B to E).

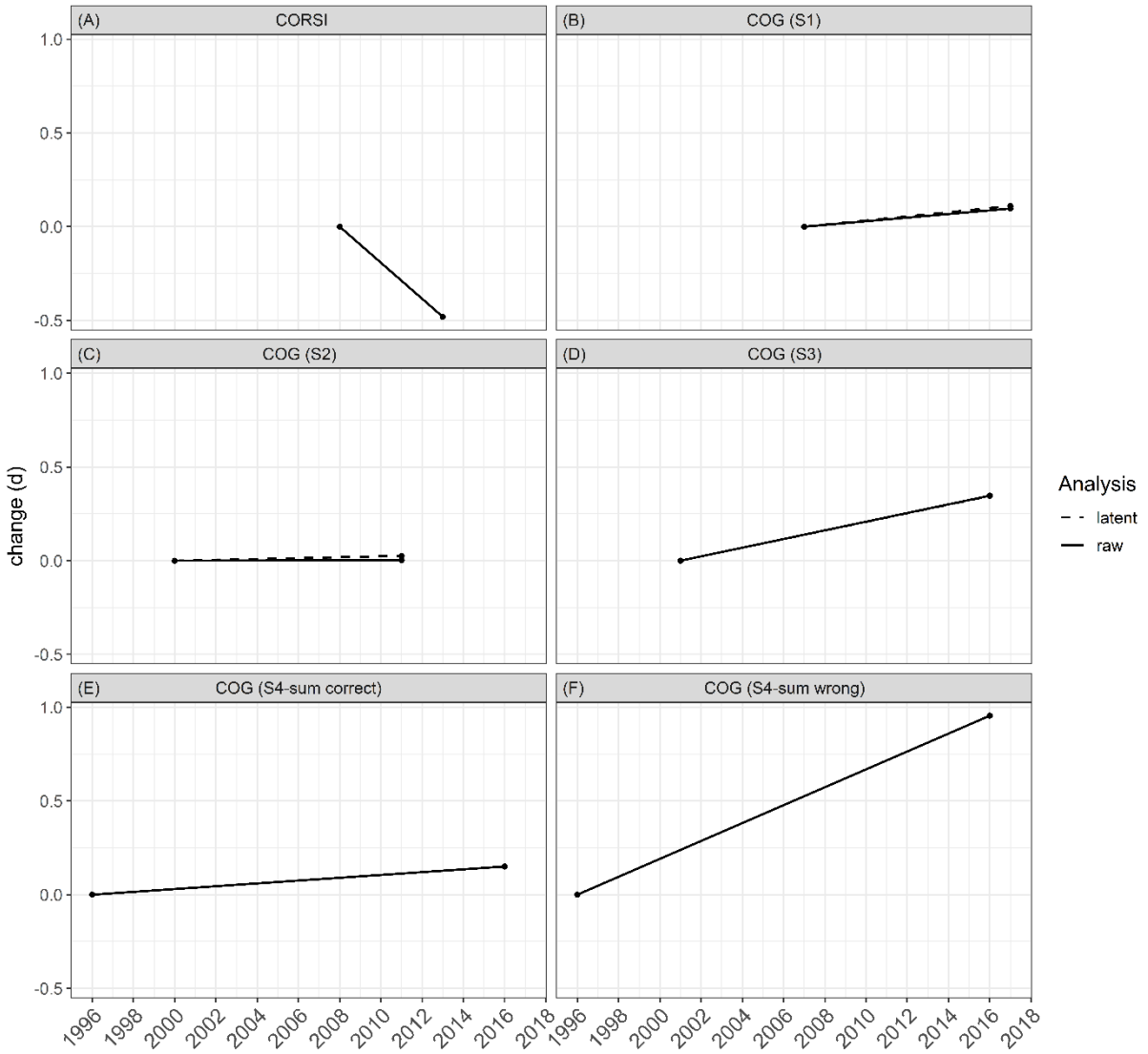

Figure 2: Test score change trajectories for the CORSI (A) and COG (B - F).

ELST-based changes showed a consistent small-to-moderate performance increase across all three subtests from 2009 to 2017 (*d*s = 0.43 to 0.61; *p*s <.001; Figure 3, panels A to C).

We observed a significant non-trivial decrease in MIP test scores between 2002 and 2011 (*d* = -0.33; *p* < .001; Figure 3, panel D).

For the SPM, we observed virtual stagnation between the two first standardization samples followed by a significant non-trivial increase. Pairwise comparisons showed trivial changes between 2001 and 2010 (*d* < 0.01; *p* = .97) and a subsequent small performance increase (*d* = 0.32; *p* < .001) from 2010 to 2017, representing a small and significant increase in test performance over the entire observed period from 2001 to 2017 (*d* = 0.28; *p* < .001; Figure 3, panel E).

LEVE changes were small and trivial (*d* = 0.01; *p* = .90) from 2007 to 2016, (Figure *3*, panel F).

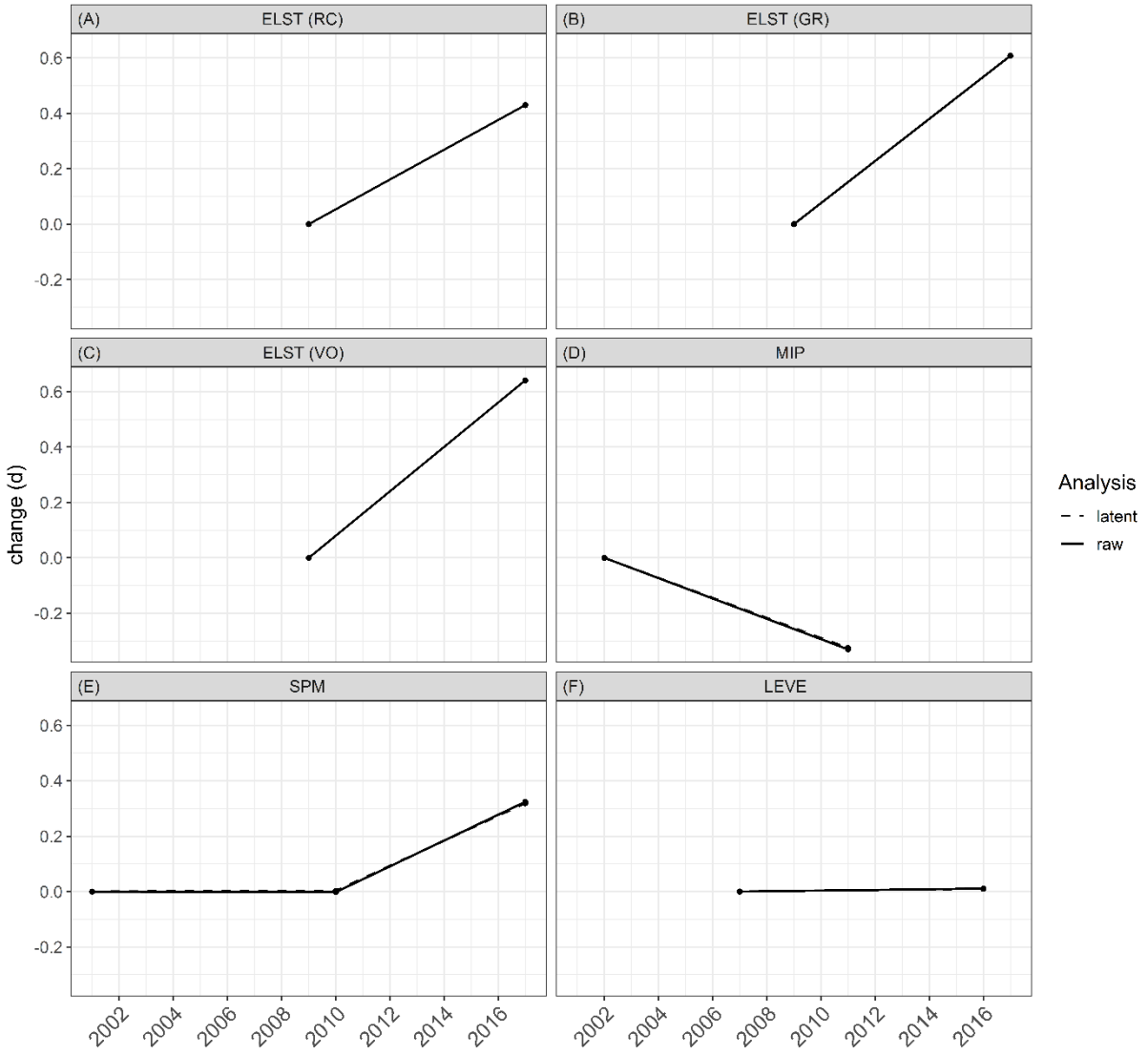

Figure 3: Test score change trajectories for the ELST (A-C), MIP (D), SPM (E), and LEVE (F).

Pairwise comparisons of RT samples, revealed a consistent performance decrease, yielding trivial-to-small effects (*d*s = -0.07 to -0.37) from 1999 to 2016. In the time interval from 1999 to 2007, we observed a trivial and non-significant decrease for the outcome variable mean reaction time (d = -0.07; p > .999), and a small significant decrease (d = -0.25; p = .003) for the mean motor time. For the subsequent time interval from 2007 to 2016, we once more observed decreases. Unlike before, performance changes in this time interval were larger for mean reaction (d = -0.37; p < .001) than mean motor time (d = -0.15; p = .15). Over the entire observed period from 1999 to 2016 this represents a moderate decrease in performance for both outcomes (mean reaction time: d = -0.44; p <.001; mean motor time: d = -0.42; p < .001; Figure 4, panels A and B).

Standardization sample-based comparisons of the 3D, revealed an erratic pattern of performance changes over time. From 2005 to 2014 performance decreases were observed (d = -0.39; p < .001) whilst performance increased from 2014 to 2016 (d = 0.29; p < .001), representing a non-significant decrease over the entire observed period from 2004 to 2016 (d = -0.11; p = .15; Figure 4, panel C).

For the 2D, we observed an initial stagnation followed by a small increase in test performance, with all effects being non-significant. Trivial performance decreases were observed from 2004 to 2010 (d = -0.07; p = .44), followed by a trivial increase in the subsequent interval from 2010 to 2016 (d = 0.14; p = .08), representing an overall increase in performance over the entire observed period from 2004 to 2016 (d = 0.07; p = .40; Figure 4, panel D).

For the ALS, group comparisons of standardization samples from 2003 and 2008 revealed a trivial average performance decrease for both the number of items that had been worked on (d = -0.07; p = .37) and error percentage (d = -0.16; p = .048; Figure *4*, panels E and F).

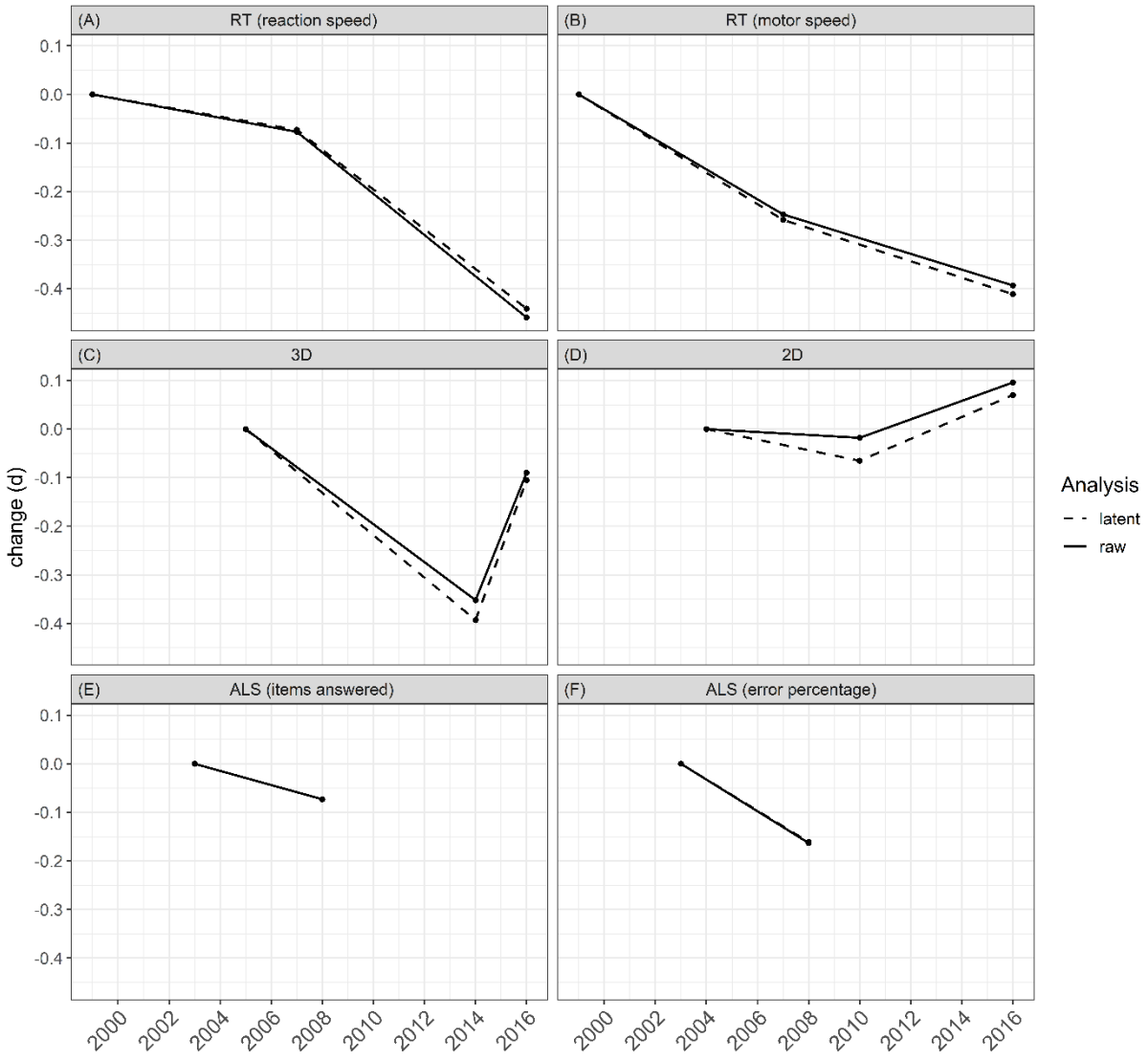

Figure 4: Test score change trajectories for the RT (A-B), 3D (C), 2D (D), and ALS (E-F).

In our supplementary analyses, we investigated potential differences in the Flynn effect for men and women, which can be assessed by examining interactions between sex and IQ test performance changes. We observed significant interactions in four subtests. Results for Visualization (2D: $\eta_p^2 = .01$; $p = .006$) showed opposing trajectories for men and women between 2004 and 2016, with performance decreasing in men, whilst increasing for women. Similarly, results from Cognitrone (COG-S1: $\eta_p^2 = .009$; $p = .005$) showed performance decreases for men and increases for women between 2007 and 2017 for test form S1, and a more rapid increase in performance for women than for men in the outcome sum of incorrect reactions in test form S4 between 1996 and 2016 (COG-S4: $\eta_p^2 = .01$; $p = .005$). For the

outcome error percentage of the Work Performance Series (ALS: $\eta_p^2 = .001$; $p = .02$) results showed performance increases in men from 2003 to 2008, but virtually no changes for women. For part A of the verbal intelligence functions subtest (NIV-A: $\eta_p^2 = .007$; $p = .01$) of the IBF, we observed that between standardization samples 2005 and 2011, performance increased more rapidly for women than for men. The small number of analyses that showed significant interactions which appeared to be unsystematic across men and women between tests and yielded exclusively trivial effects indicates that the presently observed Flynn effects were not differentiated according to sex.

Discussion

In the present study, we show an erratic pattern of IQ test score changes in stratum II domains of the CHC-model, based on a large number of largely population-representative samples of Germanophones over 22 years (1996-2018). Our results based on pairwise group comparisons of raw and latent scores derived from different and, for the most part, representative standardization samples, revealed differential temporal trajectories of IQ-test performance changes. While overall, positive changes predominated to some extent, the observed patterns vary substantially in terms of magnitude and direction, across domains, but not within domains (excepting quantitative knowledge and visual processing which both showed non-trivial effect sizes with differing signs).

Test score changes show unambiguous positive and meaningful Flynn effects for measures of full-scale cognitive ability (*g*), comprehension-knowledge (Gc), learning-efficiency (Gl), and general domain-specific knowledge (Gkn).

Processing speed (Gs) and reading and writing (Grw) did not show any meaningful test score changes whilst working memory capacity (Gwm) yielded a clear moderate negative Flynn effect.

Fluid reasoning (Gf) scores showed ambiguous results, yielding small-to-moderate increases in one, but virtual no and mixed changes in another two test instruments. This is interesting, because fluid test score gains have been typically reported to show the strongest Flynn effects when compared to other domains (Pietschnig & Voracek, 2015).

Changes in the remaining domains were ambiguous, showing meaningful effects with contrasting signs between tests in reaction and decision speed (Gt) as well as quantitative knowledge (Gq) and within tests in visual processing (Gv). This seemingly unsystematic pattern of results may reveal interesting inferences about the nature, meaning, and causes of the Flynn effect, as we discuss below.

Fluid Reasoning (Gf)

Here, we conclude that for fluid reasoning (Gf) performance gains are indicated for overlapping periods across two measures of induction (I), part A the numerical intelligence subtest (NIF-A) of the IBF and the standard version of the Raven's Progressive Matrices (SPM), between 2005 and 2018, 2010 and 2017, respectively. While the NIF-A shows clear positive changes across all three norm samples, hence the entire observed time span from 2005 to 2018, the SPM provides mixed results, with no meaningful change between 2001 and 2010, and a subsequent positive change between 2010 – 2017. The ambiguity of these results are further corroborated by standardization data from the Adaptive Matrices Test (AMT) which shows virtually no meaningful changes between 2002 and 2005, as well as between 2005 and 2014. This finding is in stark contrast with previous research, in that for markers of fluid intelligence strong gains are typically reported (e.g., Kaufmann & Lichtenberger, 2006; Pietschnig & Voracek, 2015).

Crystallized Intelligence (Gc, Grw, Gkn & Gq)

With regard to crystallized intelligence (Gc, Grw, Gkn & Gq) evidence of strong gains is provided between 2005 and 2018 by accounts of both parts of the verbal intelligence subtest (VIF-A, VIF-B) of the IBF for comprehension knowledge (Gc) and the second part (NIF-B) of the numerical intelligence subtest of the IBF for quantitative knowledge (Gq). Moreover, accounts of the English Language Skills Test (ELST) provide further evidence of strong gains in general (domain-specific) knowledge (Gkn) between 2009 and 2017, while the Reading Comprehension Test (LEVE) showed for a similar time period (2007 – 2016) stagnating effects. While on the one hand, this finding is in contrast with previous research, which generally shows lower gains in this domain, it is consistent with previous findings in German-speaking countries, providing evidence for gains, similar in size across crystallized and fluid intelligence (Pietschnig, Voracek & Formann, 2010).

Spatial Intelligence (Gv)

For measures associated with spatial domains, two markers of visual processing (Gv) and the narrow ability visualization (Vz), namely the visualization subtest of the IBF (VIZ) and the Spatial orientation test (3D), show evidence of substantial positive Flynn-effects between 2011 and 2018, 2014 and 2016, respectively. This is further supported by a relatively small performance increase from the Visualization test (2D) between 2010 and 2016.

However, all three test show negative Flynn-effects in the preceding norm comparisons (VIZ: 2005 – 2011 ; 3D: 2005 – 2014 ; 2D: 2004 – 2010). This trajectory stands in contrast with previous findings indicating a reversal of the Flynn-effect in German-speaking countries for spatial perception (Pietschnig & Gittler, 2015).

Memory (Gl & Gwm)

Measures associated with memory show differential patterns across two different cognitive abilities. Comparison of standardization samples from the Block-Tapping Test (CORSI), a marker of working memory capacity (Gwm), provides evidence for a substantial negative Flynn-effect between 2008 and 2013. Partially within the same time period, the long-term memory subtest of the IBF (LTM) shows a substantial positive Flynn-effect between 2011 and 2018. Our results for working memory capacity (Gwm) and more specifically visual-spatial short-term storage, hence are indicative of a reversal of the Flynn-effect. This stands to some extent, in contrast with previous findings reporting small positive gains for the forward Block span of CORSI over a time period of 43 years (Wongupparaj et al., 2020). Results for learning efficiency (Gl) on the other hand indicate a substantial positive Flynn-effect. Previous findings appear to be rather unclear in terms of the relation of memory and the Flynn-effect, with some reports showing no changes in performance in either direction on numerical memory tasks (Gignac, 2015), some evidence of increases in performance on visual memory tasks (Baxendale, 2010) and decreases in performance (Graves et al., 2021), and stagnation (Baxendale, 2010) on memory tasks involving verbal material.

General Speed (Gt & Gs)

We conclude that with regard to general speed, contrasting trajectories are evident across and, in part, within two different domains. Evidence from the Work Performance Series (ALS) as a measure of processing speed (Gs) and the narrow ability number facility indicates a negative Flynn-effect between 2003 and 2008. Results of two markers of reaction and decision speed (Gt), namely the Reaction Test (RT) and the Cognitrone (COG) are somewhat ambiguous. While accounts of the RT show a consistent negative Flynn-effect between 1999 and 2016, accounts of norm comparisons from COG, show a virtual stagnation for two test forms (S1: 2007 – 2017; S2: 2000 – 2011) and small-to-large positive Flynn-effects for two other test forms (S3: 2001 – 2016; S4: 1996 – 2016).

Causes and meaning of the observed changes

Several candidate theories have been proposed to account for both positive and negative Flynn effects (for an overview, see Pietschnig & Voracek, 2015; Pietschnig, Voracek, & Gittler, 2018). Based on the positive changes in knowledge-specific domains, it seems likely that education has played a considerable role in test score gains. Education quality and duration have long been suspected as meaningful contributors to test norm changes (Williams, 1998) and have been consistently supported as contributing factors for crystallized Flynn effects (Pietschnig & Voracek, 2015). However, the role of education has often deemed to be subordinate, perhaps owing to the lower strength of crystallized (i.e., more education-related) compared to fluid IQ gains. This idea is contrasted by the observation that more education-related stratum II domains showed the largest and most consistent positive test score changes whilst data from other domains showed equivocal results. Both stagnating as well as negative Flynn effects were observed in domains that had so far only been rarely investigated or had yielded inconsistent results, thus warranting replication.

It is interesting, that most of the seemingly fluid/reasoning-related subdomains yielded most of the ambiguous results. This may indicate a larger volatility of fluid and reasoning-related Flynn effects (Gt, Gq, Gv) in contrast to more knowledge-related measures.

In this vein, some particularly interesting inferences can be drawn from our finding for fluid reasoning: First, it could be argued that even though it is a Raven's-typed matrices test, test-specific properties of the AMT may have been responsible for the failure to observe any meaningful IQ changes in our study. However, we did not observe meaningful changes in the SPM from 2001 to 2010 either and only a small effect in the subsequent interval from 2010 to 2017. Because the SPM has frequently been shown to yield strong positive Flynn effects (e.g., Wongupparaj et al., 2015), this indicates that test-specific properties are not responsible for these unexpected results.

Second, it could be argued that non-representativity of the assessed samples may be the cause for these inconsistencies. However, even though three of the included samples in two of the tests did not entirely match the population census in terms of age and sex, changes in the majority of the observed intervals were based on population-representative samples and showed substantial variations between trajectories. This indicates that sample-specific characteristics are unlikely causes of these inconsistencies.

Third, Germanophone populations may genuinely differ in regard to the Flynn effect from other populations. This makes sense, because differences in IQ trajectories between countries have been well-documented and pertain to differences in regard to (i) overall changes as well as (ii) certain domains (e.g., Pietschnig et al., 2010). Despite accumulating

evidence of similarly unexpected Flynn effect trajectories in other countries (e.g., Dutton et al., 2016; Dworak, 2019; Schroeder, 2019), we cannot rule out that the Flynn effect in the Germanophone population may work in a different manner as elsewhere. Even if this were so, this does not provide an explanation for the apparent changes in the Flynn effect trajectory within the Germanophone population, which has shown substantial positive gains over most of the 1900s (Pietschnig & Voracek, 2015), but now seems to have considerably lost strength in fluid and reasoning domains.

Fourth, these differences can be interpreted as expressions of the non-linearity of the Flynn effect. Even though Flynn effect results are typically reported as linear decadal changes (as is done here as well), this is due to the typical design properties of Flynn effect studies, which do not allow for a continuous assessment. Therefore, differences between individual intervals in test score changes on the same test could be attributed to genuine changes of the Flynn effect strength or direction due to different influences (e.g., educational reforms; see Baker et al., 2015). However, non-linearity is unsuitable to explain why different tests that measure the same construct yield vastly differing effect trajectories over virtually identical time-spans.

We cannot rule out that a combination of the four above causes (e.g., non-linearity working between and test-specific properties working within cohorts) may be the reason for the observed pattern of erratic results. However, the ever-accumulating evidence of inconsistent test score changes within and between countries in the available literature seems to point towards an increasing volatility of the Flynn effect, which may be indicative of an impending stagnation or even a reversal of the Flynn effect.

Limitations

With regard to interpreting presently reported results, we need to acknowledge that observation periods were in certain cases relatively short. While we were able to include relatively long time intervals for many domains, some were separated by relatively brief intervals: Reaction and decision speed (Gt) (15 – 20 yrs), Fluid Reasoning (Gf) (12 – 16 yrs), Visual processing (Gv) (9 – 13 yrs), Quantitative knowledge (Gq) (9 – 13 yrs), Learning efficiency (Gl) (13 yrs), Comprehension-knowledge (Gc) (13 yrs), Reading and writing (Grw) (9 yrs), General (domain-specific) knowledge (Gkn) (8 yrs), Working memory capacity (Gwm) (5 yrs), Processing speed (5yrs). We see the shortest time intervals observed for measures of processing speed (Gs) and working memory capacity (Gwm), both of which provide accounts of IQ changes for period of 5 years. Similarly, general (domain-specific)

knowledge (Gkn) and reading and writing (Grw) provide observation periods of 8, 9 years respectively. This can be explained by the circumstance that for each of these four domains, only a single subtest could be included. According to common practice in Flynn effect studies, we report IQ changes per decade (DIQ), which were in some instances implausibly large in either direction. Particularly in comparisons of cohorts which were separated by comparatively brief intervals, such large absolute DIQ values were observed. These DIQs should not be interpreted at face value, because they are a function of the observed time-span, meaning that comparatively small variations between samples will result in rather large DIQs, even when effects are small. Consequently, effect size interpretations should be preferred over an interpretation of DIQs.

Another issue pertains to the representativeness of a part of the included samples. We were able to include 29 standardization samples that were representative of the respective time period in terms of age and sex, hence another 7 samples were included that were shown not to be representative (AMT: 2002, 2005; IBF: 2005; MIP: 2002; RT: 1999, 2007, 2016). Moreover, all 36 included standardization samples were not representative of the respective time period in terms of education level, as this variable was not included in the initial stratification plans. We therefore caution against interpreting these results at face value, especially with regard to possible explanations and probable causes of IQ changes. On the other hand, there is no reason to suspect that the recruiting strategy of the test publisher has led to the recruitment of samples that were systematically different in regard to their highest educational qualification over the years, beyond possible changes that are due to changes in the population education level. This interpretation is supported by the largely similar median and modal educational values between samples (see Table 2).

Final words

Our findings in this study are based on analyses of 36 Germanophone standardization samples between 1996 and 2018, the majority of which are population-representative for age and sex. Data samples from 12 well-established psychometric tests, 17 subtests and their measurements, respectively, account for a broad range of cognitive abilities and include almost all Stratum-II domains of the CHC-Framework. Pairwise group comparisons and multigroup confirmatory factor analysis (MGCFA) provide raw and latent-score changes across cohorts. This comprehensive account of IQ-changes in 10 intelligence domains contributes to disentangling the Flynn-effect in the German-speaking population, and the general population over the past two decades.

Overall, our findings reveal a rather erratic pattern of IQ-trajectories over time with accounts of a Flynn effect in Germanophones, that has become increasingly volatile, thus conforming to the recently accumulating evidence of inconsistent IQ test score changes in different countries. Our results indicate positive Flynn-effects for full scale cognitive ability (*g*), comprehension-knowledge (Gc), learning-efficiency (Gl), and general domain-specific knowledge (Gkn), a reversal of the Flynn-effect in working memory capacity (Gwm), a stagnating Flynn effect in reading and writing (Grw) and processing speed (Gs), and ambiguous effects for fluid reasoning (Gf), reaction and decision speed (Gt), quantitative knowledge (Gq), and visual processing (Gv).

Differing trajectories and strengths of changes in 10 stratum II domains of the CHC-model are consistent with previously observed negative associations between the Flynn effect and psychometric *g* (Pietschnig & Voracek, 2015).

With respect to future research we argue that it appears particularly necessary to corroborate our findings here, and supplement the literature on IQ changes in general, with further comprehensive accounts that provide insight into differential IQ changes across a wide spectrum of intelligence domains and in population-representative samples across different populations.

References

Arendasy, M. E., & Sommer, M. (2010). Evaluating the contribution of different item features to the effect size of the gender difference in three-dimensional mental rotation using automatic item generation. *Intelligence*, *38*, 574-581.

Baddeley, A. D. (2001). Is working memory still working? American Psychologist, 56, 849–864.

Baddeley, A. D. (2003). Working memory: looking back and looking forward. Nature Reviews Neuroscience, 4, 829 – 839.

Baker, D. P., Eslinger, P. J., Benavides, M., Peters, E., Dieckmann, N. F., & Leon, J. (2015). The cognitive impact of the education revolution: A possible cause of the Flynn Effect on population IQ. *Intelligence*, *49*, 144-158.

Baxendale, S. (2010). The Flynn effect and memory function. *Journal of Clinical and Experimental Neuropsychology*, *32*, 699-703.

Blum, F., Didi, H.-J., Fay, E., Maichle, U., Trost, G., Wahlen, H.-J., & Gittler, G. (2005). Basic intelligence functions (IBF) [Manual]. Mödling, Austria: Schuhfried.

Bratfisch, O., Hagman, E., & Egle, J., (2018a). Mathematics in Practice (MIP) [Manual]. Mödling, Austria: Schuhfried.

Bratfisch, O., Hagman, E., & Egle, J., (2018b). Spatial Orientation (3D) [Manual]. Mödling, Austria: Schuhfried.

Bratfisch, O., Hagman, E., & Egle, J., (2018c). Visualization (2D) [Manual]. Mödling, Austria: Schuhfried.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Corsi, P. M. (1972). Human memory and the medial temporal region of the brain. *Dissertation Abstracts International*, 34, 819B.

Cotton, S. M., Kiely, P. M., Crewther, D. P., Thomson, B., Laycock, R., & Crewther, S. G. (2005). A normative and reliability study for the Raven's Coloured Progressive Matrices for primary school aged children from Victoria, Australia. *Personality and Individual Differences*, *39*, 647-659.

European Commission. (2022). Population on 1 January by age group and sex (demo_pjangroup) [Data file]. Available from: http://epp.eurostat.ec.europa.eu/portal/page/portal/population/data/database

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., ... & Urbánek, T. (2012). Testing practices in the 21st century: Developments and European psychologists' opinions. *European Psychologist*, *17*, 300.

Dodonova Y. A., & Dodonov, Y. S. (2013). Is there any evidence of historical slowing of reaction time? No, unless we compare apples and oranges. Intelligence, *41*, 674-687.

Dutton, E., van der Linden, D., & Lynn, R. (2016). The negative Flynn Effect: A systematic literature review. *Intelligence*, *59*, 163-169.

Dworak, E. (2019). Looking for a Flynn effect: Examining shifts in cognitive ability within the SAPA project. *Oral presentation at the 20th Annual Conference of the International Society for Intelligence Research, 11.-13.07.19*, Minneapolis, MN.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, *95*, 29-51.

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.

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*, 615-632.

Guilford, J. P., & Zimmerman, W. S. (1948). The Guilford-Zimmerman Aptitude Survey. *Journal of applied Psychology*, *32*(1), 24.

Hebb, D. O. (1961). Distinctive features of learning in the higher animal. *Brain mechanisms and learning*, 37-46.

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, *32*, 175-191.

Hornke, L.F., Etzel, S., & Rettig, K. (1999). Adaptive Matrices Test (AMT) [Manual]. Mödling, Austria: Schuhfried.

Lohman, D. F. (1979). Spatial ability: A review and reanalysis of the correlational literature.

Lohman, D. F. (1989). Spatial Abilities as Traits, Processes, and Knowledge. In J. R. Sternberg, (Ed.), *Advances in the Psychology of Human Intelligence* (Vol. 4, pp. 181-248). New York: Erlbaum, Lawrence Associates.

Ortner, T., Janous, G., & Lick, E., (2020). English Language Skills Test (ELST) [Manual]. Mödling, Austria: Schuhfried.

Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: accounts of mental rotation and individual differences in spatial ability. *Psychological Review, 92*, 137.

Kaufman, A. S., & Lichtenberger, E. O. (2005). *Assessing adolescent and adult intelligence*. Hoboken, NJ: Wiley.

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modelling*, 4th ed. New York: Guilford.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*, 361-388.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*, 1-10.

Must, O., & Must, A. (2018). Speed and the Flynn effect. *Intelligence, 68*, 37-37.

Nettelbeck, T. J. (2014). Smarter but slower? A comment on Woodley, te Nijenhuis and Murphy (2013). *Intelligence, 42*, 1-4.

Nettelbeck, T., & Wilson, C. (2004). The Flynn effect: Smarter not faster. *Intelligence, 32*, 85-93.

Parker, S. (2014). Were the Victorians cleverer than us? Maybe, maybe not. *Intelligence, 47*, 1-2.

Pauli, R. (1921). Untersuchungen zur Methode des fortlaufenden Addierens. *Zeitschrift für Angewandte Psychologie*, 29, 172-187.

Peñaherrera-Aguirre, M., Fernandes, H. B., & Figueredo, A. J. (2018). What causes the anti-Flynn effect? A data synthesis and analysis of predictors. *Evolutionary Behavioral Sciences*, *12*, 276.

Pietschnig, J., Deimann, P., Hirschmann, N., & Kastner-Koller, U. (2021). The Flynn effect in Germanophone preschoolers (1996-2018): Small effects, erratic directions, and questionable interpretations. *Intelligence, 86*, 101544.

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.

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.

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*, 791-801.

Pietschnig, J., Voracek, M., & Formann, A. K. (2010). Pervasiveness of the IQ rise: A cross-temporal meta-analysis. *PLOS ONE, 5*, e14406.

Pietschnig, J., Voracek, M., & Formann, A. K. (2011). Female Flynn effects: No sex differences in generational IQ gains. *Personality and Individual Differences, 50*, 759-762.

Pietschnig, J., Voracek, M., & Gittler, G. (2018). Is the Flynn effect related to migration? Meta-analytical evidence for causes of stagnation and reversal of generational IQ test score changes. *Politische Psychologie/Journal of Political Psychology, 6*, 267-283.

Proyer, R. T., Wagner-Menghin, M. M., & Grafinger, G., (2018). Reading Comprehension Test (LEVE) [Manual]. Mödling, Austria: Schuhfried.

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 (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Raven, J., Raven, J. C., & Court, J. H., (2018). Standard Progressive Matrices (SPM) [Manual]. Mödling, Austria: Schuhfried.

Raven, J. C., & Court, J. H., (1938). *Raven's progressive matrices*. Los Angeles, CA: Western Psychological Services.

Raven, J. C., & Court, J. H., (1998). *Raven's progressive matrices and vocabulary scales* (Vol. 759). Oxford: Oxford pyschologists Press.

Rindermann, H., & Thompson, J. (2013). Ability rise in NAEP and narrowing ethnic gaps? *Intelligence, 41*, 821-831.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1-36.

Schellig, D., (2020). Block-Tapping Test (CORSI) [Manual]. Mödling, Austria: Schuhfried.

Schmidt-Atzert, L., Büttner, B. & Bühner, M. (2004). Theoretische Aspekte von Aufmerksamkeits-/Konzentrationsdiagnostik. In G. Büttner & L. Schmidt-Atzert (Hrsg.), Diagnostik von Konzentration und Aufmerksamkeit (S. 3–22). Göttingen: Hogrefe.

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* (pp. 73-163). New Haven, CT: Guilford.

Schroeder, D. (2019). A negative Flynn effect in recent cognitive ability scores. *Oral presentation at the 20th Annual Conference of the International Society for Intelligence Research, 11.-13.07.19*, Minneapolis, MN.

Schuhfried, G., (2018). Work Performance Series (ALS) [Manual]. Mödling, Austria: Schuhfried.

Schuhfried, G., (2020a). Cognitrone (COG) [Manual]. Mödling, Austria: Schuhfried.

Schuhfried, G., (2020b). Reaction Test (RT) [Manual]. Mödling, Austria: Schuhfried.

Silverman, I. W. (2010). Simple reaction time: It is not what it used to be. *The American Journal of Psychology*, *123*, 39-50.

te Nijenhuis, J., & van der Flier, H. (2013). Is the Flynn effect on g?: A meta-analysis. *Intelligence*, *41*, 802-807.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago, IL: University of Chicago Press.

Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs*.

Thurstone, L. L. (1950). Some primary abilities in visual thinking. *Proceedings of the American Philosophical Society*, *94*, 517-521.

Trahan, L. H., Stuebing, K. K., Fletcher, J. M., & Hiscock, M. (2014). The Flynn effect: a meta-analysis. *Psychological Bulletin*, *140*, 1332.

Van de Vijver, F. J. R. (1997). Meta-analysis of cross-cultural comparisons of cognitive test performance. *Journal of Cross-Cultural Psychology*, 28, 678–709.

Wechsler, D. (1981). WAIS-R [Manual]. New York: Psychological Corporation.

Westhoff, K. & Hagemeister, C. (2005). *Konzentrationsdiagnostik*. Lengerich, Germany: Pabst.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., Van Baal, G. C. M., Boomsma, D. I., & Span, M. M. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*, 509-537.

Williams, R. L. (2013). Overview of the Flynn effect. *Intelligence*, *41*, 753-764.

Williams, W. M. (1998). Are we raising smarter children today? School- and home-related influences on IQ. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 125-154). Washington, DC: American Psychological Association.

Wongupparaj, P., Kumari, V., & Morris, R. G. (2015). A cross-temporal meta-analysis of Raven's progressive matrices: Age group and developing versus developed countries. *Intelligence, 49*, 1-9.

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.

Woodley, M. A., te Nijenhuis, J., & Murphy, R. (2013). Were the Victorians cleverer than us? The decline in general intelligence estimated from a meta-analysis of the slowing of simple reaction time. *Intelligence, 41*, 843-850.

Woods, D. L., Wyma, J. M., Yund, E. W., Herron, T. J., & Reed, B. (2015). Factors influencing the latency of simple reaction time. *Frontiers in Human Neuroscience, 9*, 131.

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika, 81*, 1014-1045.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research, and Evaluation, 12*, 3.

Figures

Tables

Appendix A

Abstract

Generational IQ test score changes (the Flynn effect) were globally positive over large parts of the 20th century. However, accumulating evidence of recent studies show a rather inconsistent pattern of the Flynn effect in past decades. Patterns of recently observed test score changes appeared to be markedly different in strength and even signs between countries and domains. Because of between-study design differences and data availability in terms of differing IQ domains, so far it is unclear if these inconsistencies represent a consequence of differences in Flynn effect trajectories between countries, covered time-spans, or investigated IQ domains. Here, we present data from 36 largely population-representative Germanophone standardization samples from 12 well-established psychometric tests from 10 stratum II IQ domains from 1996 to 2018, thus providing a comprehensive assessments of domain-specific changes according to the Cattell-Horn-Carroll model of intelligence. Examination of both raw score and measurement-invariant latent mean changes yielded an erratic pattern of results, indicating positive (comprehension-knowledge, learning-efficiency, domain-specific knowledge), negative (working memory capacity), stagnating (processing speed, reading and writing), and ambiguous (fluid reasoning, reaction and decision speed, quantitative knowledge, visual processing) Flynn effects on stratum II domains. Our results show that the Flynn effect in Germanophones has become increasingly volatile over the past three decades, showing erratic patterns of test score changes between and within domains, even when limiting interpretations to measurement invariant and population representative assessments. The present findings may be indicative of an impending stagnation of the Flynn effect.

Zusammenfassung

Generationsübergreifende Veränderungen von IQ-Testergebnissen (Flynn-Effekt) waren über weite Teile des 20. Jahrhunderts weitestgehend positiv. Die sich häufenden Belege aus neueren Studien liefern jedoch vermehrt Hinweise auf ein Muster inkonsistenter Verläufe des Flynn-Effekts in den vergangenen Jahrzehnten. Diese Verläufe scheinen sich in ihrer Stärke und sogar in ihren Vorzeichen zwischen den Ländern und Bereichen deutlich zu unterscheiden. Aufgrund von Unterschieden im Studiendesign und in der Datenverfügbarkeit in Bezug auf verschiedene IQ-Domänen ist bisher unklar, ob diese Unstimmigkeiten eine Folge von Unterschieden im Verlauf des Flynn-Effekts zwischen Ländern, zwischen erfassten

Zeitspannen oder zwischen untersuchten IQ-Domänen sind. In der vorliegenden Studie präsentieren wir Ergebnisse aus 36 weitestgehend bevölkerungsrepräsentativen deutschsprachigen Normierungsstichproben aus den Jahren 1996 bis 2018 von 12 etablierten psychometrischen Tests und 10 Stratum-II-IQ-Domänen und liefern damit eine umfassende Einschätzung domänenspezifischer Veränderungen von IQ-Testergebnissen nach dem Cattell-Horn-Carroll Intelligenzmodell. Die Untersuchung sowohl der Rohwerte, als auch der messungsinvarianten latenten Mittelwertveränderungen, ergab ein inkonsistentes Muster, das über die Stratum-II-Domänen hinweg auf positive (comprehension-knowledge, learning-efficiency, domain-specific knowledge), negative (working memory capacity), stagnierende (processing speed, reading and writing) und uneindeutige (fluid reasoning, reaction and decision speed, quantitative knowledge, visual processing) Flynn-Effekte hindeutet. Unsere Ergebnisse zeigen, dass der Flynn-Effekt bei Deutschsprachigen in den letzten drei Jahrzehnten zunehmend unbeständiger geworden ist und erratische Muster von Veränderungen der Testergebnisse zwischen und innerhalb von Domänen aufweist, selbst wenn man die Interpretationen auf messungsinvariante und bevölkerungsrepräsentative Bewertungen beschränkt. Die vorliegenden Ergebnisse könnten ein Hinweis auf eine bevorstehende Stagnation des Flynn-Effekts sein**.**

Appendix B

R Syntax for MGCFA with Dichotomous Indicators

```
##########################################################################
## The present code illustrates the algorithm based on the items of all four subtests     ##
## of the Basic Intelligence Functions (IBF) test-battery, all other dichotomous          ##
## MGCFA subscale calculations were based on an identical approach                        ##
##########################################################################

### PACKAGES
for(n in c('lavaan', 'dplyr','foreign', 'effsize','parameters')){
  require(n, character.only = TRUE )}

### DATA
data = read.spss("path to datafile/datafile.sav ", to.data.frame = TRUE)

### MODEL SPECIFICATION
model <- '
g =~ verbal_A + verbal_B + spatial + memory + numerical_A + numerical_B
verbal_A =~
BEW01001+BEW01002+BEW01003+BEW01004+BEW01005+BEW01006+BEW01007+B
EW01008+BEW01009+BEW01010+BEW01011+BEW01012+BEW01013+BEW01014+BE
W01015+BEW01016
verbal_B =~
BEW02001+BEW02002+BEW02003+BEW02004+BEW02005+BEW02006+BEW02007+B
```

```
EW02008+BEW02009+BEW02010+BEW02011+BEW02012+BEW02013+BEW02014+BE
W02015+BEW02016+BEW02017+BEW02018+BEW02019

spatial =~
BEW04001+BEW04002+BEW04003+BEW04004+BEW04005+BEW04006+BEW04007+B
EW04008+BEW04009+BEW04010+BEW04011+BEW04012+BEW04013+BEW04014+BE
W04015+BEW04016+BEW04017

memory =~
BEW05001+BEW05002+BEW05003+BEW05004+BEW05005+BEW05006+BEW05007+B
EW05008+BEW05009+BEW05010+BEW05011+BEW05012+BEW05013+BEW05014+BE
W05015+BEW05016+BEW05017+BEW05018+BEW05019+BEW05020

numerical_A =~
BEW06001+BEW06002+BEW06003+BEW06004+BEW06005+BEW06006+BEW06007+B
EW06008+BEW06009+BEW06010+BEW06011+BEW06012+BEW06013+BEW06014+BE
W06015+BEW06016+BEW06017+BEW06018+BEW06019+BEW06020

numerical_B =~
BEW07001+BEW07002+BEW07003+BEW07004+BEW07005+BEW07006+BEW07007+B
EW07008+BEW07009+BEW07010+BEW07011+BEW07012+BEW07013+BEW07014+BE
W07015+BEW07016+BEW07017+BEW07018+BEW07019+BEW07020
'

### MODEL FIT FOR OVERALL DATA

overall.fit <- cfa(model,
                   data = data,
                   estimator="wlsmv",
                   std.lv = TRUE,
                   ordered = TRUE)

summary(overall.fit, standardized = TRUE, fit.measures=TRUE)


### MGCFA: model fitting for nested groups
```

```
### baseline: configural model = thresholds & loadings constrained

#for binary data, the configural or baseline model is defined by a model with constrained
#threshold and loading parameters, so that in a first step we constrain them across groups to
#see whether this model still fits within the acceptable range. If the fit of this model is not
#considerably worse than the fit of the model for the overall data then at least configural
#invariance can be assumed. We can expect the fit to drop to some extent compared to the
#overall model as the sample sizes get smaller.

config_model.fit <- cfa(model,
                        data = data,
                        ordered = TRUE,
                        meanstructure = TRUE,
                        group = "sample",
                        group.equal=c("thresholds","loadings"),
                        parameterization = "theta",
                        estimator = "wlsmv")


### residual model = strict invariance

residual_model.fit <- cfa(model,
                        data = data,
                        ordered = TRUE,
                        meanstructure = TRUE,
                        group = "sample",
                        group.equal=c("thresholds","loadings","residuals"),
                        parameterization = "theta",
                        estimator = "wlsmv")
```

```
### estimated latent (factor) means

### get parameter estimates

par = parameterEstimates(residual_model.fit)

par1 = subset(par, select = c(rhs, group, est), (group=="1" & op == "=~"))

par1_VIA = par1[grep(col_ind_VI_A, par1$rhs),]

par1_VIB = par1[grep(col_ind_VI_B, par1$rhs),]

par1_RV = par1[grep(col_ind_RV, par1$rhs),]

par1_LZG = par1[grep(col_ind_LZG, par1$rhs),]

par1_NIA = par1[grep(col_ind_NI_A, par1$rhs),]

par1_NIB = par1[grep(col_ind_NI_B, par1$rhs),]


par2 = subset(par, select = c(rhs, group, est), (group=="2" & op == "=~"))

par2_VIA = par2[grep(col_ind_VI_A, par2$rhs),]

par2_VIB = par2[grep(col_ind_VI_B, par2$rhs),]

par2_RV = par2[grep(col_ind_RV, par2$rhs),]

par2_LZG = par2[grep(col_ind_LZG, par2$rhs),]

par2_NIA = par2[grep(col_ind_NI_A, par2$rhs),]

par2_NIB = par2[grep(col_ind_NI_B, par2$rhs),]


# calculate latent continuous scores

latentScores1_VIA = apply(norm_2011[, names(items_VI_A)], 1,
function(x){x*par1_VIA$est})

latentScores2_VIA = apply(norm_2018[, names(items_VI_A)], 1,
function(x){x*par2_VIA$est})

latentScores1_VIB = apply(norm_2011[, names(items_VI_B)], 1,
function(x){x*par1_VIB$est})
```

```
latentScores2_VIB = apply(norm_2018[, names(items_VI_B)], 1,
function(x){x*par2_VIB$est})

latentScores1_RV = apply(norm_2011[, names(items_RV)], 1, function(x){x*par1_RV$est})

latentScores2_RV = apply(norm_2018[, names(items_RV)], 1, function(x){x*par2_RV$est})

latentScores1_LZG = apply(norm_2011[, names(items_LZG)], 1,
function(x){x*par1_LZG$est})

latentScores2_LZG = apply(norm_2018[, names(items_LZG)], 1,
function(x){x*par2_LZG$est})

latentScores1_NIA = apply(norm_2011[, names(items_NI_A)], 1,
function(x){x*par1_NIA$est})

latentScores2_NIA = apply(norm_2018[, names(items_NI_A)], 1,
function(x){x*par2_NIA$est})

latentScores1_NIB = apply(norm_2011[, names(items_NI_B)], 1,
function(x){x*par1_NIB$est})

latentScores2_NIB = apply(norm_2018[, names(items_NI_B)], 1,
function(x){x*par2_NIB$est})

# transpose and convert to dataframe

latentScores1_VIA = as.data.frame(t(latentScores1_VIA))

latentScores2_VIA = as.data.frame(t(latentScores2_VIA))

latentScores1_VIB = as.data.frame(t(latentScores1_VIB))

latentScores2_VIB = as.data.frame(t(latentScores2_VIB))

latentScores1_RV = as.data.frame(t(latentScores1_RV))

latentScores2_RV = as.data.frame(t(latentScores2_RV))
```

```
latentScores1_LZG = as.data.frame(t(latentScores1_LZG))

latentScores2_LZG = as.data.frame(t(latentScores2_LZG))

latentScores1_NIA = as.data.frame(t(latentScores1_NIA))

latentScores2_NIA = as.data.frame(t(latentScores2_NIA))

latentScores1_NIB = as.data.frame(t(latentScores1_NIB))

latentScores2_NIB = as.data.frame(t(latentScores2_NIB))


### calculate sumscores

latentScores1_VIA$VIA = rowSums(latentScores1_VIA, na.rm = TRUE)

latentScores1_VIB$VIB = rowSums(latentScores1_VIB, na.rm = TRUE)

latentScores1_RV$RV = rowSums(latentScores1_RV, na.rm = TRUE)

latentScores1_LZG$LZG = rowSums(latentScores1_LZG, na.rm = TRUE)

latentScores1_NIA$NIA = rowSums(latentScores1_NIA, na.rm = TRUE)

latentScores1_NIB$NIB = rowSums(latentScores1_NIB, na.rm = TRUE)


######################


latentScores2_VIA$VIA = rowSums(latentScores2_VIA, na.rm = TRUE)

latentScores2_VIB$VIB = rowSums(latentScores2_VIB, na.rm = TRUE)

latentScores2_RV$RV = rowSums(latentScores2_RV, na.rm = TRUE)

latentScores2_LZG$LZG = rowSums(latentScores2_LZG, na.rm = TRUE)

latentScores2_NIA$NIA = rowSums(latentScores2_NIA, na.rm = TRUE)

latentScores2_NIB$NIB = rowSums(latentScores2_NIB, na.rm = TRUE)


### assign group variable
```

```
latentScores1_NIB$sample = "2011"

latentScores2_NIB$sample = "2018"


### combine dataframes

latentScores1 = cbind(latentScores1_VIA, latentScores1_VIB, latentScores1_RV,
latentScores1_LZG, latentScores1_NIA, latentScores1_NIB)

latentScores2 = cbind(latentScores2_VIA, latentScores2_VIB, latentScores2_RV,
latentScores2_LZG, latentScores2_NIA, latentScores2_NIB)


##### Global Score INT

parINT1 = subset(par, select = c(lhs, rhs, group, est), (group=="1" & op == "=~" & lhs ==
"g"))

latentScores1$g = (latentScores1$VIA*parINT1$est[1] +

  latentScores1$VIB*parINT1$est[2] +

  latentScores1$RV*parINT1$est[3] +

  latentScores1$LZG*parINT1$est[4] +

  latentScores1$NIA*parINT1$est[5] +

  latentScores1$NIB*parINT1$est[6])


parINT2 = subset(par, select = c(lhs, rhs, group, est), (group=="2" & op == "=~" & lhs ==
"g"))

latentScores2$g = (latentScores2$VIA*parINT2$est[1] +

  latentScores2$VIB*parINT2$est[2] +

  latentScores2$RV*parINT2$est[3] +

  latentScores2$LZG*parINT2$est[4] +

  latentScores2$NIA*parINT2$est[5] +
```

```
  latentScores2$NIB*parINT2$est[6])


latentScores = rbind(latentScores1, latentScores2)

latentScores$sample = as.factor(latentScores$sample)


### calculate latent means

latentMeans = latentScores %>%

  dplyr::group_by(sample) %>%

  dplyr::summarise(latent_INT = mean(g), sd_INT = sd(g),

                   latent_VIA = mean(VIA), sd_VIA = sd(VIA),

                   latent_VIB = mean(VIB), sd_VIB = sd(VIB),

                   latent_RV = mean(RV), sd_RV = sd(RV),

                   latent_LZG = mean(LZG), sd_LZG = sd(LZG),

                   latent_NIA = mean(NIA), sd_NIA = sd(NIA),

                   latent_NIB = mean(NIB), sd_NIB = sd(NIB))
```

Appendix C

R Syntax for MGCFA with Continuous Indicators

```
##########################################################################
## The present code illustrates the algorithm based on the items of all four subtests     ##
## of the Reaction Test (RT), all other continuous MGCFA subscale                         ##
## calculations were based on an identical approach                                       ##
##########################################################################


### PACKAGES
for(n in c('lavaan', 'dplyr','foreign', 'effsize','parameters')){
  require(n, character.only = TRUE )}


### DATA
data = read.spss("path to datafile/datafile.sav ", to.data.frame = TRUE)


### MODEL SPECIFICATION
model <- '
reactionSpeed =~
RZEIT001+RZEIT007+RZEIT013+RZEIT016+RZEIT018+RZEIT019+RZEIT022+RZEIT0
24+RZEIT025+RZEIT031+RZEIT037+RZEIT040+RZEIT042+RZEIT043+RZEIT046+RZE
IT048
motorSpeed =~
MZEIT001+MZEIT007+MZEIT013+MZEIT016+MZEIT018+MZEIT019+MZEIT022+MZ
EIT024+MZEIT025+MZEIT031+MZEIT037+MZEIT040+MZEIT042+MZEIT043+MZEIT
046+MZEIT048
```

```
'

### model fit for the overall data
overall.fit <- cfa(model,
                   data = data,
                   std.lv = TRUE,
                   estimator = "MLM")

summary(overall.fit, standardized = TRUE, fit.measures=TRUE)

### model fitting for nested groups
```

In the following section the model will be fitted groupwise, stepwise inducing more restrictions by imposing constraints across the groups.

```
### configural model
configural_model.fit <- cfa(model,
                            data = data,
                            estimator="MLM",
                            std.lv = TRUE,
                            meanstructure = TRUE,
                            group = "sample")

summary(configural_model.fit, standardized = TRUE, fit.measures=TRUE)

### metric model
```

```
metric_model.fit <- cfa(model,
                        data = data,
                        estimator="MLM",
                        std.lv = TRUE,
                        group = "sample",
                        meanstructure = TRUE,
                        group.equal = "loadings")

summary(metric_model.fit, standardized = TRUE, fit.measures=TRUE)

#### scalar model

scalar_model.fit <- cfa(model,
                        data = data,
                        estimator="MLM",
                        std.lv = TRUE,
                        meanstructure = TRUE,
                        group = "sample",
                        group.equal = c("loadings", "intercepts"))

summary(scalar_model.fit, standardized = TRUE, fit.measures=TRUE)

#### residual model
residual_model.fit <- cfa(model,
                          data = data,
```

```
                estimator="MLM",
                std.lv = TRUE,
                group = "sample",
                meanstructure = TRUE,
                group.equal = c("loadings", "intercepts", "residuals"))

summary(residual_model.fit, standardized = TRUE, fit.measures=TRUE)

### estimated latent (factor) means
# get parameter estimates

par = parameterEstimates(residual_model.fit)

par1 = subset(par, select = c(rhs, group, est), (group=="1" & op == "=~"))
par1_MRZ = par1[grep(col_ind_score_MRZ, par1$rhs),]
par1_MMZ = par1[grep(col_ind_score_MMZ, par1$rhs),]

par2 = subset(par, select = c(rhs, group, est), (group=="2" & op == "=~"))
par2_MRZ = par2[grep(col_ind_score_MRZ, par2$rhs),]
par2_MMZ = par2[grep(col_ind_score_MMZ, par2$rhs),]

par3 = subset(par, select = c(rhs, group, est), (group=="3" & op == "=~"))
par3_MRZ = par3[grep(col_ind_score_MRZ, par3$rhs),]
par3_MMZ = par3[grep(col_ind_score_MMZ, par3$rhs),]
# calculate latent continuous scores
```

```
latentScores1_MRZ = apply(norm_1999[, names(items_MRZ)], 1,
function(x){x*par1_MRZ$est})

latentScores2_MRZ = apply(norm_2007[, names(items_MRZ)], 1,
function(x){x*par2_MRZ$est})

latentScores3_MRZ = apply(norm_2016[, names(items_MRZ)], 1,
function(x){x*par3_MRZ$est})


latentScores1_MMZ = apply(norm_1999[, names(items_MMZ)], 1,
function(x){x*par1_MMZ$est})

latentScores2_MMZ = apply(norm_2007[, names(items_MMZ)], 1,
function(x){x*par2_MMZ$est})

latentScores3_MMZ = apply(norm_2016[, names(items_MMZ)], 1,
function(x){x*par3_MMZ$est})


# transpose and convert to dataframe

latentScores1_MRZ = as.data.frame(t(latentScores1_MRZ))

latentScores2_MRZ = as.data.frame(t(latentScores2_MRZ))

latentScores3_MRZ = as.data.frame(t(latentScores3_MRZ))


latentScores1_MMZ = as.data.frame(t(latentScores1_MMZ))

latentScores2_MMZ = as.data.frame(t(latentScores2_MMZ))

latentScores3_MMZ = as.data.frame(t(latentScores3_MMZ))


# calculate sumscores

latentScores1_MRZ$MRZ = rowMeans(latentScores1_MRZ, na.rm = TRUE)

latentScores2_MRZ$MRZ = rowMeans(latentScores2_MRZ, na.rm = TRUE)
```

```
latentScores3_MRZ$MRZ = rowMeans(latentScores3_MRZ, na.rm = TRUE)


latentScores1_MMZ$MMZ = rowMeans(latentScores1_MMZ, na.rm = TRUE)
latentScores2_MMZ$MMZ = rowMeans(latentScores2_MMZ, na.rm = TRUE)
latentScores3_MMZ$MMZ = rowMeans(latentScores3_MMZ, na.rm = TRUE)


# assign group variable
latentScores1_MRZ$sample = "1999"
latentScores2_MRZ$sample = "2007"
latentScores3_MRZ$sample = "2016"


#combine dataframes
latentScores1 = cbind(latentScores1_MRZ, latentScores1_MMZ)
latentScores2 = cbind(latentScores2_MRZ, latentScores2_MMZ)
latentScores3 = cbind(latentScores3_MRZ, latentScores3_MMZ)


latentScores = rbind(latentScores1, latentScores2, latentScores3)
latentScores$sample = as.factor(latentScores$sample)


# calculate latent means
latentMeans = latentScores %>%
  dplyr::group_by(sample) %>%
  dplyr::summarise(latent_MRZ = mean(MRZ), sd_MRZ = sd(MRZ),
                   latent_MMZ = mean(MMZ), sd_MMZ = sd(MMZ))
```

Appendix D

Table C1

Model summaries for measurement invariance analyses

Basic Intelligence Functions (IBF)

Model	Chi-Square	df	RMSEA	SRMR	CFI	ChangeCFI
Configural Invariance	12620.109	12300.000	0.008	0.095	0.999	0.000
Strict Invariance	14160.438	12412.000	0.019	0.100	0.993	-0.006

Cognitrone (GOG) - test form S1

Model	Chi-Square	df	RMSEA	SRMR	CFI	ChangeCFI
Configural Invariance	714.397	340.000	0.051	0.034	0.943	-0.002
Weak Invariance	709.620	359.000	0.048	0.049	0.945	0.002
Strong Invariance	742.297	378.000	0.048	0.049	0.945	0.000
Strict Invariance	749.761	398.000	0.460	0.046	0.945	0.000

Cognitrone (GOG) - test form S2

Model	Chi-Square	df	RMSEA	SRMR	CFI	ChangeCFI
Configural Invariance	629.851	340.000	0.057	0.047	0.930	-0.004
Weak Invariance	650.231	359.000	0.055	0.060	0.931	0.001
Strong Invariance	711.120	378.000	0.058	0.062	0.924	-0.007
Strict Invariance	580.512	398.000	0.042	0.142	0.916	-0.008

Cognitrone (GOG) - test form S3

Model	Chi-Square	df	RMSEA	SRMR	CFI	ChangeCFI
Configural Invariance	2785.484	1804.000	0.045	0.042	0.871	-0.031
Weak Invariance	2740.066	1847.000	0.043	0.066	0.878	0.007
Strong Invariance	2837.698	1890.000	0.043	0.067	0.873	-0.005
Strict Invariance	2424.479	1934.000	0.031	0.075	0.911	0.038

Cognitrone (GOG) - test form S4

Model	Chi-Square	df	RMSEA	SRMR	CFI	ChangeCFI
Configural Invariance	5582.530	1624.000	0.078	0.101	0.987	-0.003
Strict Invariance	-	-	-	-	-	-

Mathematics in Practice (MIP)

Model	Chi-Square	df	RMSEA	SRMR	CFI	ChangeCFI
Configural Invariance	373.054	358.000	0.012	0.117	0.996	-0.003
Strict Invariance	423.910	378.000	0.021	0.125	0.988	-0.008

Raven's Standard Progressive Matrices (SPM)

Model	Chi-Square	df	RMSEA	SRMR	CFI	ChangeCFI
Configural Invariance	73.484	93.000	0.000	0.024	1.000	0.000
Strict Invariance	94.135	103.000	0.000	0.028	1.000	0.000

Reaction Test (RT)

Model	Chi-Square	df	RMSEA	SRMR	CFI	ChangeCFI
Configural Invariance	2249.126	1389.000	0.042	0.047	0.945	-0.006
Weak Invariance	2302.272	1449.000	0.041	0.056	0.945	0.000
Strong Invariance	2410.844	1509.000	0.041	0.057	0.943	-0.002
Strict Invariance	2144.457	1573.000	0.032	0.069	0.946	0.003

Reading Comprehension Test (LEVE)

Model	Chi-Square	df	RMSEA	SRMR	CFI	ChangeCFI
Configural Invariance	274.441	222.000	0.026	0.103	0.962	0.009
Strict Invariance	283.642	238.000	0.023	0.105	0.967	0.005

Spatial Orientation (3D)

Model	Chi-Square	df	RMSEA	SRMR	CFI	ChangeCFI
Configural Invariance	2807.625	871.000	0.079	0.180	0.962	-0.007
Partial Strict Invariance	3112.676	921.000	0.082	0.186	0.957	-0.005

Visualization (2D)

Model	Chi-Square	df	RMSEA	SRMR	CFI	ChangeCFI
Configural Invariance	1737.662	650.000	0.074	0.176	0.916	-0.060
Strict Invariance	1895.546	694.000	0.075	0.185	0.907	0.009

Work Performance Series (ALS)

Model	Chi-Square	df	RMSEA	SRMR	CFI	ChangeCFI
Configural Invariance	2991.160	1478.000	0.058	0.048	0.939	0.006

Weak Invariance	3062.854	1516.000	0.058	0.054	0.938	-0.001
Strong Invariance	3106.131	1554.000	0.057	0.057	0.938	0.000
Strict Invariance	3292.189	1594.000	0.059	0.059	0.932	-0.006

Appendix E

Declaration of authorship credit

Substantial parts of this work, originate from, and were in part published verbatim, in Lazaridis, Vetter, & Pietschnig (https://doi.org/10.1016/j.intell.2022.101707).

As co-author, I made a significant contribution to the article and share responsibility and accountability for all research findings published therein. Personal contributions to the article comprise the majority of the transcription including introduction and theory, methods, and results. Further personal contributions include all aspects of the data collection, data preparation, and data analysis. These contributions can therefore be found to a large extent and in part, in verbatim, in the present work.