MASTERARBEIT | MASTER'S THESIS

Titel | Title

Measurement-invariant IQ trajectories – temporal changes in fluid intelligence (2001-2024)

verfasst von | submitted by

Moritz Fuchs BSc

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

Master of Science (MSc)

Wien | Vienna, 2025

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
Betreut von | Supervisor:	Assoz. Prof. Mag. Dr. Jakob Pietschnig Privatdoz.

Abstract

Positive generational IQ test score changes (i.e., the Flynn effect) have been globally observed for large parts of the 20th century. More recent research from the past decades, however, revealed quite ambiguous patterns, regarding strength and even direction of IQ trajectories across different countries and intelligence domains. Specifically, in the domain of fluid intelligence where, historically, gains have been most robust and largest, inconsistent findings seemingly point to a deceleration, stagnation or even a reversal of the Flynn effect. Here, we present measurement invariant data from four population-representative Germanophone cohorts' (total *N* = 1223) performance in an inductive reasoning task (Raven's SPM; Raven, et al., 2018) across four measurement points (2001, 2011, 2018 and 2024). Analyses revealed non-linear patterns. After performance stagnated between 2001 and 2011, positive changes could be documented subsequently. While some incremental changes between cohorts amounted to 11 IQ points per decade, the total detected Flynn effect over the observed time period from 2001 to 2024 amounted to changes of approximately 4 IQ points per decade, corresponding to a medium effect (Cohen *d* = 0.61). Our findings emphasize the importance of establishing measurement invariance in order to interpret Flynn effects properly. The implications of our results are being discussed in light of recently observed trends of domain-specific IQ trends and mechanisms of ability differentiation.

Introduction

Intelligence: Definition and theoretical foundation

To this day, despite broad conceptual consensus on the matter, there is no verbatim definition of intelligence. The one closest to being unanimously accepted is the following by Gottfredson (1997): ‘Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience’ (p. 13). Accordingly, intelligence is not merely a narrow skill but rather reflects an individual’s broader and deeper capability for comprehending its environment, enabling him to ‘make sense’ of things (Gottfredson, 1997). Speaking in more psychometric terms, most intelligence experts use *psychometric g* as a working definition of intelligence (Gottfredson & Saklofske, 2009). Psychometric *g* was first postulated by Spearman (1904) who realized that students’ performances in different school subjects are positively intercorrelated. From this observation he derived the hypothesis of a common, underlying mental ability which transcends different domains and special contexts – a general factor of intelligence (*g-factor*).

It has been debated whether psychometric *g* is merely a statistical artifact or, in fact, a biologically manifest factor. Advocating for the latter, Jensen (1999) interpreted the *g* factor as a biological, genetically determined phenomenon. He argued that *g* serves as a foundation to human cognition and can be thought of analogously in terms of physical laws. Thereby, *g* can not only be measured by psychometric tests but also through biological indicators. In line with his research, many authors found associations of neurophysiological correlates with *g*, ranging from neuronal conduction velocity, reaction times and brain size (Jensen, 1999) to cortical excitability (Eysenck, 1988), efficiency of glucose metabolism and amount of grey matter in the prefrontal cortex (Haier et al., 1988).

Other authors disagree with the idea of *g* being biologically determined by arguing that the g-factor is merely a statistical abstraction and not a ‘thing’ in the brain (Gould, 1981). Thereby, positive correlations between test scores do not prove a common biological foundation, as they may stem from overlapping cognitive processes or environmental influences. Flynn (1987, 2012) also rejected the notion that IQ can be viewed as synonymous to innate, biologically determined intelligence, highlighting the environmental and sociocultural malleability of intelligence. Alternative intelligence models like Gardner's (1993) multiple intelligences framework reject the concept of *g* altogether, as it describes complex mental abilities as a network of several autonomous, culturally embedded intelligences with distinct evolutionary developments. However, subsequent investigations contradict the idea of multiple autonomous intelligences, as they correlate under another significantly (Visser et al., 2006a). Furthermore, despite Gardner's efforts to neglect the influence of *g*, each of these multiple intelligences are evidently related to *g* (Visser et al., 2006b), suggesting that Gardner's framework is merely another (though inaccurate) form of expressing psychometric *g*. Further critique involves the fact that these eight domains are of striking similarity to the second stratum factors in modern intelligence models (Carroll, 1993). In consideration of the multitude of biological correlates (see Visser et al., 2006b), it is safe to say that *g* is at least partially biologically determined. However, definitive conclusions about biological differences and their implications in *g* between populations should be drawn with caution in light of evolutionary theories (e.g., Lynn, 2006) about racial differences in intelligence (Wicherts et al., 2010).

Building on Spearman's *g*-factor theory, a distinction of psychometric *g* in two broad facets of intelligence was developed (Cattell, 1963; Horn & Cattell, 1966): *fluid intelligence* (G_f), which reflects biologically determined, culture-independent processes of mental processing and solving new problems, and *crystallized intelligence* (G_c), which captures abilities that are explicitly or implicitly learned across a lifetime. These learning processes are characterized by a transformation of fluid abilities into crystallized capacities. The G_f – G_c model gained substantial empirical support and laid the conceptual groundwork for modern, more nuanced models of intelligence, such as the *Cattell–Horn–Carroll (CHC) model*, which integrates updated versions of G_f (the ability to reason, solve novel problems, and identify patterns independent of prior learning) and G_c (knowledge and skills acquired through education and cultural experience) alongside other intelligence domains within a unified framework (Schneider & McGrew, 2018).

Initial research on generational IQ changes

Early findings from intelligence research have soon detected that IQ scores seem to vary over time. Initially, cross-sectional studies have generally found a decrease in cognitive performance after a peak in the early twenties or thirties (e.g., Horn & Cattell, 1966; Jones & Conrad, 1933; Schaie, 1958). Rising test scores – even though noticed – were mainly attributed to statistical artifacts, sampling errors or to intrapersonal developmental changes (Merrill, 1938; Tuddenham, 1948). Being the first to do so, Schaie and Strother (1968) eventually interpreted age-related cognitive changes as generational cohort differences. These findings, subsequently documented by James Flynn (1984) laid the foundation of what became known as the *Flynn effect*. Initially, Flynn (1987) drew emphasis on the distinction between genetically fixed cognitive potential and environmentally malleable IQ test performance. He argued that differences between generations in IQ scores do not reflect a quantitative increase in intelligence but rather can be viewed as qualitative shifts in cognitive adaptation shaped by differing environmental demands. Accordingly, IQ tests do not measure 'real' intelligence but rather socially prioritized abilities with their measured scores being subject to fluctuation (Flynn, 1987). During the 1990s he began to modify his view on this regard. He acknowledged that gains – instead of being mere test artifacts – reflect genuine improvements of culturally relevant and -dependent skills (Flynn, 1999). Flynn (2012) attributed these improvements to what he called *scientific spectacles* – a specific way of thinking mediated by modern societies' education system prioritizing abstract classification, logical analysis and hypothetical problem solving. Thereby, rising scores indicate real cognitive advances in specific domains without being a manifestation of a uniform increase in general intelligence (Flynn, 2012).

Building on this shift, scholars have been repeatedly emphasized the societal relevance of the Flynn effect (e.g., Kanaya et al., 2003). A particular controversial instance regarding US death penalty underscores this matter (Flynn, 2012). As the Supreme Court exempts individuals with an IQ below 70 (i.e., threshold to learning disability) from the death penalty, inconsistent interpretations of Flynn effect trends may lead to the execution of individuals who would in reality qualify for having an intellectual disability, leading to what Flynn (2012) describes as '[…] a lottery in which life and death [that] are decided by what test they happened to take […]' (p. 66). Hence, accurately accounting for the impact of the Flynn effect extends beyond an academic discourse and has far reaching sociopolitical relevance for our society.

Global IQ gains

Empirical evidence for the Flynn effect

Many contemporary studies support the notion of a positive Flynn effect across many intelligence domains. Ranging from 2-4 IQ points per decade (DIQ) several large-scale meta-analyses confirmed IQ-gains found by individual studies in numerous countries all over the world (e.g., Pietschnig & Voracek, 2015; Trahan et al., 2014). In China, cohort-dependent upsurges in both vocabulary and mathematical abilities amounted to DIQ = 3.2 and DIQ = 3.8, respectively (Qi & Xiong, 2023). Similar gains were found by a meta-analysis in Romania, detecting full-scale IQ gains of DIQ = 3.4 across many age groups in the population (Gunnesch-Luca & Iliescu, 2020). Other findings from Europe align well with these proposed margins (e.g., Spain: DIQ = 3.5, Colom et al., 2023; France: DIQ = 3.0, Bradmetz & Mathy, 2006; Austria, Germany and Suisse: DIQ = 3.5, Pietschnig et al., 2010). A recent example for exceptionally large Flynn effects stems in Austria, with changes in several intelligence domains (e.g., fluid reasoning, visual processing, quantitative knowledge) ranging from 8.6 to 14.8 IQ points per decade (Andrzejewski et al., 2024). It has to be noted that the authors recommend interpreting these test score changes not numerically, but merely directionally. Similar IQ increases were prevalent in English children with changes amounting up to 12 IQ points per decade for fluid intelligence (Lynn et al., 1987). IQ increases in these amounts are unusual for western, developed countries, but do mirror findings from eastern developing nations. Research on the Flynn effect in Sudan has yielded large IQ gains of DIQ = 8.4 (Dutton et al., 2018) and DIQ = 7.2 (Elbanna et al., 2018) for fluid intelligence. Similarly, Qatar and Libya could demonstrate equivalent Flynn effects in the same intelligence domain, amounting to DIQ = 7.6 (Warrag et al., 2018) and DIQ = 9.4 (Al-Shahomee et al., 2018), respectively. Such different expressions of IQ gains from different geographic regions all over the world highlight the global relevance and variability of generational IQ trajectories.

Factors influencing the Flynn effect: age, ability and socioeconomic context

It appears that IQ gains do not follow uniform patterns. Instead, they seem to be influenced by a multitude of sociodemographic and contextual variables, such as age, cognitive ability levels or national wealth.

Regarding the moderating influence of age on the Flynn effect, no clear consensus is present. Some studies report stronger gains in younger generations (Mertens et al., 2022), others in adults (Pietschnig & Voracek, 2015) while some find no age differences at all (Trahan et al., 2014). Differences in their findings may be partly explained by the analysis of different IQ tests, measuring different intelligence domains (full scale IQ: Trahan et al., 2014; full scale IQ, fluid IQ, crystallized IQ & spatial IQ: Pietschnig & Voracek, 2015;), as well as differing geographical research scopes (USA & Great Britain: Trahan et al., 2014; Global: Pietschnig & Voracek, 2015).

Another ongoing debate concerns the topic of whether the Flynn effect occurs independently from cognitive ability levels. Some researchers show that people with higher as well as with lower intelligence both have exhibited similar improvements in their test performance as a result of generational score increases (Pietschnig & Voracek, 2015; Trahan et al., 2014). Others show contrasting results by stating that ability increases seem to be of a greater magnitude in lower ability segments compared to upper segments (Oberleiter et al., 2025) as well as being solely present in lower IQ regions (Flynn & Shayer, 2018). Such increases in performance of individuals from the lower tail of the IQ distribution has been proposed to be a significant factor for generational test score gains in recent years (Pietschnig et al., 2013). Thereby, the Flynn effect may be associated with a reduction of the population variability instead of a holistic shift of the IQ distribution (Rodgers, 1998; see Oberleiter et al., 2025). Alternatively, contradicting results suggest that Flynn effects were only present for IQs > 130 and IQ decreases just for IQs < 70, reflecting possible inequalities regarding accessibility to Flynn effect-enhancing environments (Platt et al., 2019). However, due to the particularity that their data stems from a study that examines the prevalence and comorbidities of psychiatric disorders, the results are not generalizable.

Beyond age- and ability variations, it has been theorized that the magnitude of IQ gains depends on national wealth. IQ scores have been found to be higher in developed countries, but IQ increases to be larger in low-income, emerging countries (Brouwers et al., 2009). Others disagree, suggesting that gains in full-scale, crystallized and spatial IQ are stronger in developed nations, whereas gains in fluid IQ were equally present in developing as well as developed countries (Pietschnig & Voracek, 2015). This is interesting as both findings seemingly appear to be intuitively plausible, even though they suggest contrasting results. Larger Flynn effects in developed countries are congruent with the idea that rising IQ scores are associated with better educational systems or higher living standards (as discussed later; see Rindermann et al., 2017). On the other hand, developing countries might have a greater potential to increase their IQ scores as opposed to developed countries who potentially face ceiling effects (see Meisenberg & Lynn, 2023). While such categorizations whether a nation is regarded as developed or not are time- and viewer-dependent, the extent to which the Flynn effect varies according to certain national economic circumstances remains debated.

Inconsistent Flynn effect-patterns

While age, ability levels and national wealth can explain varying quantities of IQ gains across sociodemographic contexts, they cannot fully account for recent inconsistencies in Flynn effect patterns that point to a deceleration or even reversal of the Flynn effect (Oberleiter et al., 2024b).

Deceleration and stagnation

Differences in growth rates is no new phenomenon. Many studies have detected IQ gains to be non-linear over time (e.g., Flynn, 2007) with earlier cohort usually recording larger gains than more recent cohorts, especially after the 1980s (Pietschnig & Voracek, 2015). However, recently observed anomalies in several countries hint at a potential stagnation of the Flynn effect. For instance, gains in different Wechsler scales have been hypothesized to reach a plateau in 2024 (Russell, 2007) – a notion supported by the examination of standardization samples from the WIAS-IV (Wechsler, 2008) who showed a decelerating Flynn effect of DIQ = 1.2 (Winter et al., 2024). Accordingly, analysis of NAEP data from thousands of students in the US showed a deceleration of IQ gains (DIQ = 1.2) between 1971 and 2008 (Rindermann & Thompson, 2013). Alongside gains varying across different subjects (e.g., gains in reading were lower than in mathematics), a narrowing of ethnic gaps has been detected in regard to their achievement scores, pointing to differing growth rates for different sociocultural contexts.

Several findings from other countries match these results found in the US. For example, average IQ gains over several IQ domains in Spanish students have decreased significantly (DIQ = 1.7, Colom et al., 2023). Similar patterns were observed in Australian primary school students (Cotton et al., 2005) as well as in Germanophone preschoolers in Austria and Germany (Pietschnig et al., 2021).

Some assume that stagnating IQ trajectories might not necessarily reflect a general deceleration of the Flynn effect but rather age specific developmental dynamics. It could be assumed that the Flynn effect does not manifest until children reach certain progress in school (see Pietschnig et al., 2021). On the other hand, a global deceleration of Flynn effect patterns has been indeed documented and cannot be ruled out with certainty (Pietschnig & Voracek, 2015).

Reversal and decline

Emphasizing the ambiguous nature of Flynn effect research even further, some authors speculate that the Flynn effect has not only reached a plateau but may, in fact, be declining. While the only claim for a global decline of the Flynn effect stems from a systematic literature review consisting of a small sample size (N = 9 studies; Dutton et al., 2016), numerous empirical investigations report regional instances of decreasing cognitive test scores across different domains and age groups.

Early suggestions of a *negative Flynn effect* stem from Norway, where a continuous decline in IQ scores could be observed after 1995 (Sundet et al., 2004). A similar pattern was present in Denmark, with significant decreases of verbal, numerical and spatial test performance amounting to DIQ = -3.0 in the late 1990s and early 2000s (Teasdale & Owen, 2008). Notably, the authors ruled out the possibility of test-immanent factors like ceiling effects being responsible for these declines, potentially hinting at environmental circumstances at work. However, as both studies used data from young men liable for military service, the generalizability of their results might be reduced. Data from other Scandinavian countries (e.g., Finland: Dutton & Lynn, 2013; Koivunen, 2007; Norway: Bratsberg & Rogeberg, 2018; Sweden: Emanuelson et al., 1993) as well as Estonia (Kõrgesaar, 2013) indicate similar trends.

Similar patterns have been reported in central Europe. A meta-analytic approach from the Netherlands using reaction times as a proxy for IQ found large decreases of DIQ = -4.5 (Woodley & Meisenberg, 2013). It has to be noted, however, that due to variations in the measurement instruments, which evidently lead to a deceleration of reaction times, the validity of their results have been severely questioned (e.g., Flynn, 2013; Parker, 2014; Woods et al., 2015). Dutton & Lynn (2015) found decreases in France on a full-scale IQ level in similar amounts than previous authors. Their results, subsequently, have been shown to be attributable to measurement non-invariance, so that the reported negative Flynn effect is actually a positive one (Gonthier et al., 2022). More recently, equivalent Flynn effect patterns have been observed in Germanophone countries (mostly in Austria and Germany) as well. In addition to a decline in visual processing (DIQ = -4.8) in the three-dimensional cubes (3DC; Gittler, 1984) test for spatial perception (Pietschnig & Gittler, 2014), declines were also evident in figural reasoning as a subdomain of fluid intelligence (DIQ = -4.7) from 2012 to 2022 (Oberleiter et al., 2024b).

Broadening the scope of the field to English speaking countries, negative Flynn effect patterns have also been prevalent in Britain (e.g., Shayer & Ginsburg 2007; 2009) and the US (e.g., Dworak et al., 2023). Notably, Shayer and Ginsburg (2007; 2009) did not use common, standardized IQ tests, but Piaget's volume & heaviness- (Piaget & Inhelder, 1975), pendulum (Inhelder & Piaget, 1958), and equilibrium (Piaget, 1985) test, potentially limiting comparability of their findings.

Taken together, it appears that most of the findings on the negative Flynn effect are either methodologically flawed, take place in isolated intelligence domains, or reflect country-specific trends. Rather than pointing to a global reversal of the Flynn effect, these results seem to exemplify fundamental challenges in Flynn effect research. Fragmented data across geographical contexts and large timespans are difficult to interpret, as the Flynn effect can only be studied retrospectively. Additionally, cross-temporal comparisons are limited by temporal shifts in test properties – an issue we will return on later. However, based on the current state of Flynn effect research, a global reversal of the Flynn effect cannot be ruled out with complete certainty.

Underlying causes of the Flynn effect

These increasingly ambiguous Flynn effect patterns are accompanied by an effort to link a wide range of possible causes to observed IQ trajectories. The following section provides an overview over some widely proposed biological, environmental and hybrid causes eligible of explaining both, IQ gains and IQ losses.

Biological, genetical and demographical factors

Hybrid vigor. The *hybrid vigor* (also referred to as *heterosis*) hypothesis accounts IQ gains to humans increasingly mating with individuals from genetically dissimilar subpopulations, thereby increasing allelic heterozygosity (Pietschnig & Voracek, 2015). Among the hypothesized causes for this phenomenon, also referred to as *heterosis*, are the breakup of small, isolated communities due to urbanization, greater population mobility and, as a result, globalization (Mingroni, 2007). However, a reevaluation of this hypothesis revealed little explanatory power in regard to the magnitude of observed gains (Woodley, 2011). Even under optimal conditions (i.e., maximal outbreeding), heterosis is only able to account for a fraction of observed IQ gains amounting to 3 IQ opints over 50 years, indicating that hybrid vigor alone is not able to explain global IQ gains (Pietschnig & Voracek, 2015).

Migration. While the hybrid vigor hypothesis emphasizes the benefits of the mixing of ethnic subpopulations for intelligence, some authors describe *migration* as a possible cause for the negative Flynn effect (Woodley of Menie et al., 2018; Dutton et al., 2016). Not only has it been shown that non-Western ethnic minorities (e.g., sub-Saharan Africans, South Asians) in Western countries score lower on IQ tests than European (Lynn, 2006; Dutton et al., 2016) but it could also be demonstrated that immigration correlates positively with IQ losses (Woodley of Menie et al., 2018). As IQ losses through immigration appear to be happening predominately in culture-loaded IQ tests strongly associated with *g* (i.e., full-scale IQ, crystallized IQ) and not in low *g*-ness tests (i.e., fluid IQ, spatial IQ), differing IQs of that sort can easily be explained by immigrants' lower proficiencies in the native language of their adoptive countries (Woodley of Menie et al., 2018). Apart from immigration, emigration has been named as responsible for negative IQ trends, too. Such *'brain-drain'* effects occur, when highly educated and intelligent individuals leave a certain country to seek better living opportunities in more developed countries (Rindermann & Thompson, 2016). Consequently, a reduction of the average IQ of the remaining population is to be expected (Woodley of Menie et al., 2018).

Taken together, it appears that neither biological nor demographical causes are able to explain the Flynne effect exhaustively. Biological models, such as the hybrid vigor hypothesis, fail to account for findings on siblings who – while sharing the same parents and, thus, the same genetical prerequisites – do exhibit Flynn effects depending on their birth year (Sundet et al., 2010). Similarly, direct investigation of migration effects yielded no evidence for associations between migration and negative Flynn effect patterns (Pietschnig et al., 2018).

Environmental factors

Nutrition. An intuitive explanation for the Flynn effect which is commonly adapted by many researchers is a rise in global understanding of and access to sufficient *nutrition* (e.g., Colom et al, 2005; Colom et al., 2007; Lynn, 2009). Increased nutrition has been linked to increases in both height and brain size, with these changes being of a similar magnitude to those observed in intelligence scores (Pietschnig & Voracek, 2015). The nutrition-hypothesis, however, has been challenged by James Flynn (2012) who pointed out that even severe instances of malnutrition like the Dutch famine in 1944 were not accompanied by declines in IQ scores and could, therefore, not have had a measurable impact on cognitive performance. While nutrition certainly affects the well-being of the human body and has been linked to other, positively with IQ test performance correlated biological characteristics like head circumference (Lucas et al., 1998), the impact and mechanisms on the Flynn effect remain unclear.

Education. Another widely acknowledged explanation is the global improvements in educational systems (e.g., Rindermann et al., 2017). *Education* is strongly associated with G_c and, as a result, gains in G_c have been interpreted to enhanced schooling standards (McArdle et al., 2002). Paradoxically, achievement test scores, like SATs, have not been affected by periods of increasing IQ scores (Flynn, 1984). This contradiction, however, might be due to increasingly differing test-taker demographics, as access to better education continuously included broader and more heterogenous segments of the population (Teasdale & Owen, 1987). The impact of education on G_f is somewhat more unclear, even though a growing body of evidence suggests that increasing exposure to cognitively stimulating environments and early pre-school programs might be of supporting nature (Gorey, 2001; Suna & Ozer, 2024). Intriguingly, the impact of education appears to be somewhat ambiguous, as the very same educational improvements once driving IQ gains may now be a possible factor for decelerating growth rates in more recent years (Pietschnig & Voracek, 2015) or even the overall reversal of the Flynn effect. As educational systems improve, they may fall victim to ceiling effects and, consequently, yield diminishing benefits for cognitive improvements (see Oberleiter et al., 2025; Pietschnig et al., 2018). This could explain the fact that the countries that were among the first showing negative Flynn effect trends, like Finland (Dutton & Lynn, 2013), France (Dutton & Lynn, 2015) and Denmark (Teasdale & Owen, 2008) are among the wealthiest nations in the world where environmental IQ-enhancing factors, such as education, may have already peaked (Oberleiter et al., 2025). Other considerations comprise a rising ‘anti-intellectual culture’ (see Woodley of Menie et al., 2018) – meaning declines in educational values as well as negative changes in educational exposure and -quality – as possible causes for a reversal of the Flynn effect (Bratsberg & Rogeberg, 2018). Contrasting Flynn’s (2012) notion of modern societies exerting ‘scientific spectacles’, education has been considered to be decreasingly science-oriented while, at the same time, yielding a decline in sophisticated reading culture (Dutton & Lynn, 2015). In sum, education appears to be an important, but not sufficient factor in explaining positive, negative and stagnating Flynn effect patterns.

Pathogen stress. Similar to explanatory models that emphasize the importance of physiological inputs (e.g., the nutrition hypothesis) efforts have been made to establish a link between Flynn effect patterns and the exposure to aversive *pathogens*. Thereby, less risk of exposure to aversive pathogens has been proposed as a contributing factor to IQ gains, as fending off detrimental pathogens during early childhood occupies crucial biological resources otherwise needed for brain development (see Pietschnig & Voracek, 2015). In line with this idea, country-specific IQ scores are negatively associated with the global prevalence of infectious diseases (Eppig et al., 2010). Nevertheless, effects of resource shortages at birth due to environmental pathogen exposure can be overridden by a reduction of disease burden during pubertal development (Hõrak & Valge, 2015). While access to health care services or hygiene in total have increased globally (WHO & The World Bank, 2023), increasing air pollution in developing countries has been shown to impede cognitive performance in math- and word recognition tests (Zhang et al., 2018), yielding explanatory approaches for stagnating or declining IQ scores. It remains to be seen to what extent the COVID-19 pandemic will have an influence on cognitive performance of affected cohorts. Preliminary findings indicate that infants born during the pandemic exhibit significantly reduced verbal, nonverbal, and overall cognitive performance compared to children born pre-pandemic (e.g., Deoni, 2022). However, findings like these should be interpreted with caution as extensive longitudinal data on this topic is lacking yet.

Hybrid Factors

Social multipliers. Dickens and Flynn (2001) proposed a model that considered gene-environment interactions and their multiplying effects on one another. These *social multipliers* create a positive feedback loop, enabling even small changes in environment properties to emerge in large performance gains in a short period of time. Biological determinants and environmental variability should, thereby, be understood as complementary factors that cannot work independently. A basketball analogy helps to illustrate this mechanism (see Dickens & Flynn, 2001): Basketball performance is certainly genetically influenced (e.g., by body height) but also shaped by the environment (e.g., through training). Multiplying effects arise when new broadcasting technologies, such as television, are introduced, boosting the sports' popularity. Practitioners become more enthusiastic and, therefore, engage more in practices, enhancing their basketball ability, and so on (Dickens & Flynn, 2001). This feedback loop, however, can also work in the opposite direction, creating *negative social multiplier* effects (Woodley of Menie et al., 2018). Diminishing opportunities for cognitive development may exacerbate declining human capital and, consequently the environments necessary to sustain or promote cognitive gains, possibly leading to stagnating or even reversing Flynn effects. If, for instance, educational quality decreases, individual intellectual engagement may decline which, in turn, reinforces schooling quality (see Woodley of Menie et al., 2018; Meisenberg, 2003). While many Flynn effect-related observations do fit well in the social multiplier hypothesis (e.g., stronger gains in fluid IQ), most empirical investigations cannot provide a direct test of social multiplier effects due to the inherent difficulty of isolating complex, co-dependent interactions between genes and environment across larger time periods (Pietschnig & Voracek, 2015).

Life history speed. The *life history speed* model does not suggest novel causes for Flynn effects but attempts to integrate several proposed factors in a comprehensive framework (Pietschnig & Voracek, 2015). It postulates the idea that changes in environmental conditions (e.g., nutrition, pathogen exposure, family sizes, education) influence life history speed (Woodley, 2012). Individuals with slower life history speed are typically associated with fewer sexual partners, fewer offspring and later parenthood as compared to individuals with fast life history speed (Pietschnig & Voracek, 2015). While environments with high mortality rates and bad living conditions promote individuals with fast life history speed, safe and stable environments support individuals with slow life history speed. Consequently, sufficient available resources and enhanced living standards encourage the emergence of slow life history speed resulting in the development of differentiated cognitive abilities which have been hypothesized to be a driving factor for rising IQ scores (Woodley, 2012; Woodley et al., 2024). As the model does not presuppose uniform changes across countries or timespans and allows for compensatory effects between different life speed history enhancing and inhibiting factors, it conceptually fits well with erratic, seemingly contradictory Flynn effect patterns (Pietschnig & Voracek, 2015; Woodley et al., 2013; Woodley et al., 2024).

In summary, proposed causes of the Flynn effect contain a broad spectrum of biological, environmental and hybrid explanatory models. Proposed biological and environmental causes contribute only partially to a coherent picture as they only capture aspects of the complex interplay of mechanisms shaping generational cognitive trends. Hybrid models such as social multipliers and life-history speed give a more comprehensive framework but remain difficult to test empirically. In order to understand recently observed erratic and contradictory IQ trajectories it is, therefore, crucial to consider psychometric perspectives alongside environmental, biological or hybrid determinants.

Domain specificity of the Flynn effect

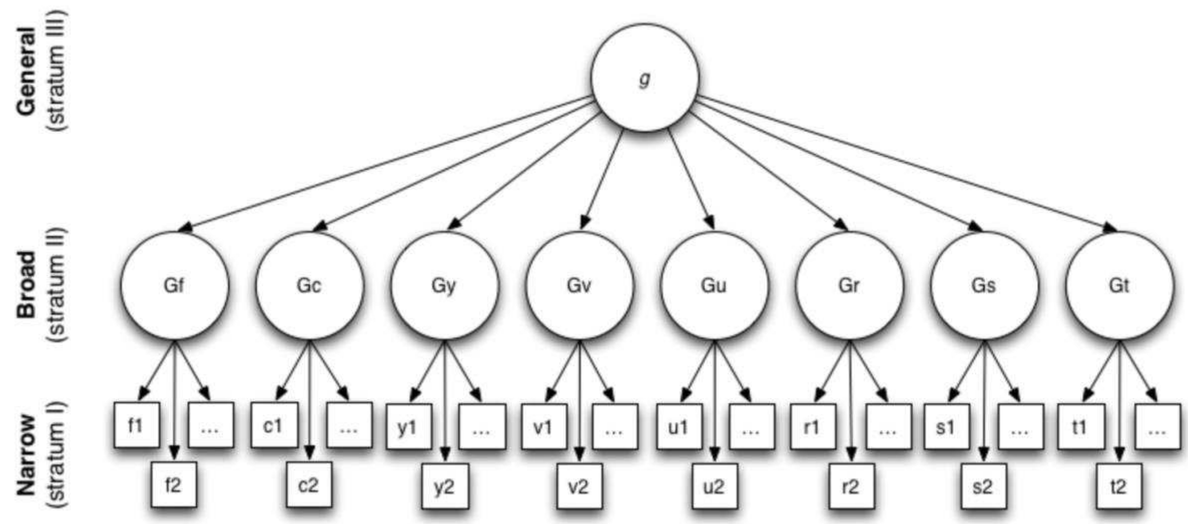
Domain specific gains in the CHC-model

While much of the discourse in Flynn effect research focusses determining whether and to what extent intergenerational IQ changes vary depending on specific environmental and biological circumstances, a growing body of research points to the importance of differentiating IQ gains depending on the specific intelligence domain in which they occur. For most of the 20th century, generational IQ gains were contextualized within Cattell and Horn’s (1966) classical dichotomy of G_f and G_c (Pietschnig et al., 2023). IQ increases between these two domains have been shown to differ with gains in G_f being about twice as big than those in G_c (Pietschnig & Voracek, 2015). Many other studies align with this notion, hinting at differing developmental rates of domain-specific IQ trajectories (Flynn, 2007; Lynn, 1990; Lynn & Hampson, 1986).

While the taxonomy between G_f and G_c initially provided a useful theoretical foundation for intelligence research, it offered only a limited perspective on the complex developmental patterns of intelligence (Oberleiter et al., 2024b). Modern, more fine-grained assessments of cognitive abilities provide a more nuanced understanding of the structure and development of intelligence. In the widely adopted Cattell-Horn-Carrol (CHC) model (Schneider & McGrew, 2018), G_f and G_c are conceptualized as broad abilities, that exist on the same level of abstraction as other intelligence domains, such as visual processing (G_v), auditory processing (G_a), processing speed (G_s), reaction and decision speed (G_t), short-term memory (G_{sm}), long-term storage and retrieval (G_{lr}), reading and writing (G_{rw}), quantitative reasoning (G_q) and general knowledge (G_{kn}) (see Lazaridis et al., 2022). Subordinated in each of these broad (Stratum II) abilities are several narrow (Stratum I) abilities, e.g. G_f includes inductive reasoning, sequential reasoning and figural reasoning (Schneider & McGrew, 2018; see Figure 1). Consequently, the CHC-model enables a more differentiated investigation of positive, stagnating or negative Flynn effect patterns on stratum II as well as on stratum I level, as generational IQ changes may vary across different (sub)domains instead of being reflected as global, uniform changes in overall intelligence (Oberleiter et al., 2024b). Evidence from measurement-invariant test score changes within Austrian population-representative samples further illustrates the domain-specific nature of the Flynn effect: Several stratum II domains exhibited differing IQ trajectories, including (a) positive (comprehension-knowledge, learning-efficiency, domain-specific knowledge), (b) negative (working memory capacity), (c) stagnating (processing speed, reading and writing), and (d) ambiguous (fluid reasoning, reaction and decision speed, quantitative knowledge, visual processing) trends, with some stratum I abilities varying directionally within the same stratum II domain (Lazaridis et al., 2022).

Figure 1

The Cattell-Horn-Carrol (CHC) intelligence model



Note. Reprinted from Bates (2013)

The Flynn effect in Raven's Standard Progressive Matrices

Interestingly, among the cognitive domains exhibiting ambiguous trajectories in Lazaridis and colleagues' (2022) findings is fluid reasoning – a domain which has consistently been yielding the most robust gains across generations (Flynn, 1998). A particularly influential measurement tool for assessing these gains in fluid reasoning is *Raven's Standard Progressive Matrices* (*SPM;* Raven, 1938). Due to its nonverbal testing format and its appropriateness for a broad age range (Raven et al., 2018), Raven's SPM is one of the most regularly used IQ tests worldwide and, historically, has been essential in defining the empirical status of the Flynn effect (Flynn, 2012; Nordmo et al., 2025).

A large body of studies report large intergenerational SPM-gains. Compared to other IQ measures in the US and UK, SPM gains were twice as large (DIQ = 7.0) compared to gains from other IQ measures (Trahan et al., 2014). Similarly, SPM specific gains (DIQ = 5.7) were substantially larger than overall gains across IQ tests in the US (DIQ = 3.0; Williams, 2013). Other industrialized countries, such as Israel, Norway, Belgium and the Netherlands exhibited similar gains of DIQ = 6.0 on average (Flynn, 1999), as well as parts of the Chinese (DIQ = 4.4; Neisser; 1998) and the Spanish (DIQ = 6.9; Colom et al., 1998) population. Findings of that sort lead some authors to characterize figure matrices tests as the main driving force behind rising IQ scores (Flynn, 2012; Nordmo et al., 2025). Interestingly, SPM-specific meta-analytic investigations yielded somewhat smaller gains of DIQ = 2.0 (Brouwers et al., 2009), falling short of other meta-analytical findings of DIQ = 4.1 for fluid intelligence (Pietschnig & Voracek, 2015).

Despite its popularity and several claims about Raven's SPM showing less influence of confounding cultural factors on cross-national differences than any other IQ test (e.g., Brouwers et al., 2009), evaluations on the SPM's factorial structure and stability of measurement properties challenge such overly enthusiastic appraisals. In addition to the fact that different item sets require differing solving strategies and, therefore, do not measure the underlying construct unidimensionally (Lynn et al., 2004; Mackintosh & Bennett, 2005), it has been shown that several items assess intelligence differently depending on measurement point or geographical context (e.g., Fox & Mitchum, 2013; Wicherts et al., 2010).

In sum, Raven's SPM seem to be an influential driving factor of Flynn effect patterns, yet its psychometric issues highlight the need to critically examine these performance changes and determine whether they are real latent ability changes or test artefacts. This, in turn, raises questions of how these domain-specific change patterns relate to *g*.

The Flynn effect and psychometric *g*

The role of g-loadings in Flynn effect patterns

Throughout the history of intelligence research, it has been evident that various measures of IQ subdomains are associated with *g* in different ways. Pietschnig &

Voracek (2015) proposed a useful taxonomy categorizing certain intelligence domains in relation to their *g-ness*: high (crystallized IQ), medium (full-scale IQ) and low (fluid IQ). Accordingly, it has been shown that IQ subtests with low *g*-ness yield larger changes in test performance over time than subtest with high *g*-ness, which corresponds to the well-established finding of gains in G_f being substantially larger than in G_c (e.g., Pietschnig & Voracek, 2015; Woodley of Menie et al., 2018). A prime example for this mechanism is Raven's SPM. As a nonverbal measurement of inductive reasoning – a subdomain of fluid intelligence – increases in SPM test scores do not remarkably correspond with increases in psychometric *g* (Gignac, 2015). We can, therefore, interpret Flynn effects in low *g*-ness domains like fluid intelligence to reflect specific cognitive ability shifts, in contrast to those in high *g*-ness domains, which are considered to yield changes closer to general intelligence (te Nijenhuis & van der Flier, 2013).

Changes in the positive manifold

Consequently, generational test score changes – especially in fluid intelligence and Raven's SPM – may not express global changes in *g*, but rather reflect differentiated ability improvements on stratum II and stratum I level (Lazaridis et al., 2022; Oberleiter et al., 2024b; Oberleiter et al., 2025; Pietschnig et al., 2021; Pietschnig et al., 2023). As Flynn effects have ben predominately detected in low *g*-ness domains, a growing body of research shows a negative association of IQ gains with psychometric *g* (Andrzejewski et al., 2024; Pietschnig & Voracek, 2015; Woodley & Madison, 2013). Due to asymmetric cognitive ability trajectories in certain individual domains (i.e., some abilities increase to a greater extent than others) across the majority of the population the intercorrelations between IQ domains on stratum II level are weakened, resulting in an overall weakening of the positive manifold (Andrzejewski et al., 2024; Oberleiter et al., 2024a). This mechanism can be illustrated by an analogy to the decathlon (see Andrzejewski et al., 2024). Performances in different sub-disciplines (i.e., hurdling, pole vault, discus throw) are usually positively intercorrelated, resulting in a positive manifold of athletic performance. Therefore, athletes performing well in one discipline are likely to also perform well in other disciplines, indicating a general athleticism factor. As the achieved sum scores across all disciplines represent the overall decathlon performance, training one specific discipline (i.e., long jump) intensively will not only lead to higher scores in that very discipline but also to a better total decathlon performance. Focusing training efforts on just one discipline to a great extent, however, may over time affect performance in other disciplines outside that focal point. Interestingly, overall performance will keep on rising as long as the specific gains in the trained discipline outweigh the losses from the disciplines not trained explicitly (i.e., as long as performance in long jump is increased in a greater fashion than the losses in every other discipline combined). Due to the law of diminishing returns (Spearman, 1904), the same amount of training, however, will yield continuously smaller performance gains, up to the point where losses in all other disciplines may outweigh incremental gains in one specific discipline, leading to an overall decathlon performance decrease. As a result, the decreasing decathlon performance is not the result of a given decathlete having become a worse athlete but rather a consequence of increased ability specialization (i.e., they have become a long jumper). This, in turn, will affect his general athleticism factor in a negative way, as his long jump ability does not indicate a good performance in other disciplines anymore, leading to a weakening of the positive manifold (Andrzejewski et al., 2024).

Empirical evidence on this matter has been given by investigating variations of intercorrelations between several WISC-subtests in Japanese children (Lynn & Cooper, 1994), changes in achievement *g* in Italian students (Pietschnig et al., 2023) and decreases in the correlation of different IBF (Blum, 2018) subscales (Andrzejewski et al., 2024). In light of the independence between the Flynn effect and changes in the underlying structure of intelligence – especially regarding IQ gains in low *g*-ness domains, such as fluid intelligence – it becomes increasingly crucial to reliably and accurately differentiate between genuine ability changes and confounding measurement artifacts, such as temporal shifts in test properties.

Methodological challenges of intergenerational IQ comparisons

Measurement invariance

An essential methodology to interpret generational IQ changes accurately is the establishment of *measurement invariance* (see Meredith, 1993). Measurement invariance enables comparability of cross-temporal or longitudinal data by ensuring that (a) the underlying factorial structure is identical (*configural invariance*), (b) the factor loadings are consistent (*metric invariance*), (c) the variables' intercepts are equivalent (*scalar invariance*) and (d) the residual variances of the outcomes are equal (*strict invariance*) (van de Schoot et al., 2012). For binary IQ test items, the configural model is statistically equivalent to a model with constrained thresholds and factor loadings (Wu & Estabrook, 2016).

Differential item functioning (DIF)

If an IQ measure turns out not to be measurement invariant, its items are subjected to what is called *differential item functioning* (*DIF*). DIF occurs when differences in average performance between two cohorts are caused by differences in the items' ability to discriminate between genuine ability levels (also referred to as *item drift*; Gonthier & Gregoire, 2022). As societal norms and cultural understanding evolve, test takers engage with these items with varying levels of knowledge, which leads to variations in how difficult certain items are perceived to be (Oberleiter et al., 2024b). An example by Wicherts (2007) helps to illustrate this issue (see also Gonthier & Gregoire, 2022): People scored higher on a vocabulary test item asking for the definition of the word 'terminate' after the movie 'Terminator' was released and, conversely, scored lower on an item asking for the definition of the word 'Kremlin' after the release of the movie 'The Gremlins'. While the test scores changed in both of these instances, the latent underlying intellectual ability did not.

These considerations have important implications for intelligence research, as evolving, culturally dependent item properties can bias the interpretation of Flynn effects. A recent example of such biases involved negative Flynn effect patterns in France (see Dutton & Lynn, 2015) that could be attributed not to cognitive regression but to item drift in some Wechsler scales (e.g., WAIS, WISC; Wechsler, 1974, 1981), meaning that their measurement properties did in fact change over time (Gonthier et al., 2021; for similar phenomena see also Tran et al., 2014). Others found that widely used IQ measures, such as different Wechsler scales, did not measure cognitive performance invariantly across subpopulations (e.g., Winter et al., 2024), hinting at the fact that at least some of the historically reported Flynn effect patterns may have come into existence through DIF (Oberleiter et al., 2024a).

Thus, only when cross-temporal measurement invariance is established – indicating that item properties remain constant over time and no DIF is at work – can test score changes be interpreted as true ability changes rather than measurement artifacts (Lazaridis et al., 2022).

Present Study

Considering the recently observed inconsistencies in Flynn effect patterns, it is crucial to determine whether intergenerational IQ score changes reflect genuine cognitive improvements or whether they are biased by temporal shifts in measurement properties. Comparisons of different cohorts from different measurement time periods should, in addition to the establishment of measurement invariance, primarily rely on IQ tests involving as little culture-based declarative knowledge as possible (Gonthier et al., 2021). One such test is Raven's SPM, which plays a crucial role in Flynn effect research due to its global applicability across various age groups and its culture-reduced, nonverbal test design (Raven et al., 2018). These particularities make Raven's SPM less susceptible to confounding sociocultural factors and, therefore, DIF across (sub)populations (Brouwers et al., 2009).

Despite its critical role in interpreting Flynn effects across several domains, very few studies to date have critically examined whether IQ measures used to detect Flynn effects were able to uphold measurement invariance across generational cohorts or measurement points. The present study aims to address this gap by investigating whether test score changes in Raven's SPM (total N = 1,223) can be interpreted meaningfully as genuine changes in inductive reasoning abilities across four comparable cohorts. To do so, we collected data from a Germanophone sample representative of 2001's demographic composition in 2024 and compared it to three archival standardization samples from 2001, 2011 and 2018.

Methods

In our study, we investigated changes in IQ trajectories by comparing four Germanophone cohorts using Ravens Standard Progressive Matrices (SPM) over a period of 23 years. Study design, sampling rationale, analysis plan and the main study hypothesis were pre-registered on the Open Science Framework (OSF; https://osf.io/sv7cy).

Participants

We obtained three archival SPM cohorts from 2001, N = 275, M_{age} = 43.6, 64.4 % female; 2011, N = 320, M_{age} = 43.6, 54.4 % female, and 2018, N = 340, M_{age} = 49.9, 52.1 % female, that were originally collected by a renown Austrian test publisher (Schuhfried GmbH) as SPM standardization samples. All cohorts are population-representative for their respective collection year. A detailed description of all three archival cohorts including the purposefully collected sample from 2024 is provided in Table 1.

Data collection took place from April 2024 until March 2025. Prior to taking part in the study, people had to indicate (a) their age in years, (b) their sex (male or female) and (c) their level of education according to the predefined metric displayed in Table 2. To be eligible for participation, individuals had to be proficient in German, and their age had to be in the range between 16 and 85. We originally aimed to recruit a sample of N = 275 participants whose demographic profile consisting of age, sex and level of education closely mirrors that of the 2001 cohort. Due to oversampling of educational group five (University- or College degree) we ended up recruiting 13 additional people of educational group 3 (Technical College or Vocational Training) in order to correct for this imbalance resulting in a total sample size of N = 288, M_{age} = 42.8, 58.7 % female.

Table 1

Sample characteristics according to cohort

	2001	2011	2018	2024
N	275	320	340	288
Age				
Range	16-85	14-81	16-87	16-85
M	43.6	43.6	49.9	42.8
SD	17.5	16.9	17.0	19.4
Sex				
Male	98	146	163	119
Female	177	174	177	169
Education Level				
1	3	7	1	3
2	31	54	37	35
3	108	160	120	102
4	105	69	122	107
5	28	30	60	41

Table 2

Level of Education

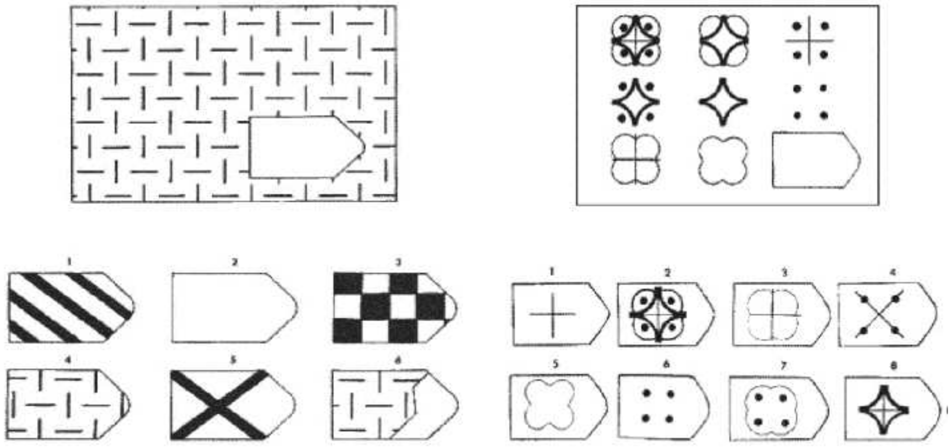
Level	Description	Years
1	No school leaving certificate or special school	< 9
2	Completed compulsory school or secondary school	9-10
3	Completed technical college or vocational training	10-12
4	Completed secondary school with Abitur/Matura	12-13
5	University or College degree	> 13

Materials

For our study, we used Raven's Standard Progressive Matrices (SPM; Raven et al., 2018); one of the most widely used nonverbal assessments of inductive reasoning abilities (Evers, 2012). Locating it in the CHC-taxonomy, the overall SPM score of participants serves as an indicator for fluid reasoning on stratum II level and inductive reasoning on stratum I level (Lazaridis et al., 2022). It is widely recognized for its robust internal consistency of α = .77-.96 and its appropriateness for a broad age range (Raven et al., 2018). The test comprises 60 figure matrices items, each of which consists of nine elements in a 3x3 matrix. The task involves identifying the underlying pattern of eight displayed elements and selecting the ninth missing element from a set of six or eight alternatives which completes the pattern. The items are grouped into five distinct sets that progressively increase in difficulty: (A) Simple patterns: Basic shapes and colors, requiring minimal abstraction; (B) Complex spatial patterns: Items with increased complexity through spatial relations; (C) Abstract logical patterns: Requires participants to identify underlying logical relations; (D) Advanced abstract patterns: Presents more demanding abstract reasoning tasks; (E) Challenging matrices: Items with multiple plausible solutions, testing higher-order reasoning (Raven et al., 2018). Two example items – an easy item from set A and a difficult item from set E – are provided in Figure 1 (Zainuddin et al., 2018).

Figure 2

Raven's SPM example items from set A (left) and set E (right)



Note. Reprinted from Zainuddin et al. (2018)

Procedure

Participants were mainly recruited via flyers on public train, bus and tram stops as well as on social media channels (i.e., Facebook groups, Reddit threads). Specific spaces and buildings (i.e., schools, churches, labor market service, retirement homes, community buildings) were either contacted via email or asked to hang out advertisement flyers in order to target specific age- and educational groups. Additionally, participants that took part in the study were encouraged to pass contacts of the research team on to their friends, family and colleagues.

In every instance – including the 2024 data collection – the test was administered via a computer in a standardized single-, or group setting (up to four participants at a time), replicating the conditions used in earlier data collections. The 2024 testing procedure took place in the Computer Assessment Lab at the Department of Differential Psychology and Psychological Diagnostics, Faculty of Psychology, University of Vienna. Prior to undergoing testing, participants received detailed information about the study procedure and the associated privacy policy. They were required to sign an informed consent, confirming their comprehension and agreement to participate. Participants were allowed to stop the test at any given moment without having to face any consequences. After being verbally informed about the procedure and properties of the IQ assessment, the test was administered with a duration of approximately 45 minutes. Besides participants being informed about their results by receiving an evaluation form; no monetary remuneration or compensations of other kind was provided.

Statistical Analysis

For statistical Analysis we used R 4.2.2 (R Core Team, 2025) and RStudio 2025.05.1+513 (R Studio Team, 2025); measurement invariance analyses were performed with the lavaan (version 6.0-18) R package (Rosseel, 2012). The analysis code as well as all used packages are provided in the Appendix.

First, pairwise standardized mean differences (Cohen *d*, 1988) between the raw item scores across all four samples (2001, 2011, 2018 and 2024) were calculated. Furthermore, ANOVAs with cohort as the independent variable and item scores as the dependent variable were performed to test for the formal significance of the changes in test scores.

Second, we conducted measurement invariance analyses in order to assess whether the items consistently measure the same latent construct across different cohorts. This was done by using multi-group confirmatory factor analyses (MGCFAs), following the approach by Oberleiter and colleagues (2024b), with four cohorts representing distinct groups in our analysis. To determine measurement invariance, we compared models of increasing restrictiveness and interpreted the resulting changes in the model-fit indices. Following the approach by Wu and Estabrook (2016), a model with invariant thresholds and loadings (configural invariance) has been used as a baseline model. To evaluate model comparisons, we used changes in Comparative Fit Indices (CFI) between two models. Changes in CFI > 0.01 would suggest non-maintenance of measurement invariance at the level of the more restrictive model (Cheung & Rensvold, 2002). If strict measurement invariance is established, we can conclude that (a) the underlying factorial structure of the latent construct measured by the measurement instrument has not changed, (b), the items' association with the latent construct remained constant, (c) item difficulty was consistent over the observed measurement points, and (d) the items' measurement errors remained equivalent across all cohorts.

Third, based on the resulting models, to investigate changes in the latent construct we estimated latent factor scores and means. Afterwards, pairwise standardized mean differences (Cohen *d*) based on these latent means were calculated and transformed into annual IQ changes per decade (DIQ) via the following formula: DIQ = [(Cohen *d* x 15) / interval in years] x 10 (see Andrzejewski et al., 2024).

Results

Our analysis revealed a positive Flynn Effect from 2011 to 2024, while a negative, though marginal and non-significant, effect was observed between 2001 and 2011. The mean SPM scores of all cohorts are shown in Table 3.

Table 3

Mean SPM scores across all cohorts

Cohort	*N*	M_{score}	SD_{score}
2001	275	47.5	10.1
2011	320	46.9	7.6
2018	340	49.4	7.7
2024	288	52.7	6.0

Analysis of variances (ANOVAs) showed a significant main effect of cohort, $df = 3$, $F = 31.62$, $p < .001$, $\eta^2 = 0.07$, indicating that 7% of variance in IQ scores can be explained by cohort differences, representing a medium effect size. Pairwise comparisons yielded a non-significant difference between 2001 and 2011 in test scores. However, from 2011 onwards, test scores increased significantly over the analyzed timespan, with proportions of explained variance ranging from $\eta^2 = 0.01$ to $\eta^2 = 0.15$. Detailed results are presented in Table 4.

Before conducting measurement invariance analysis, we first had to exclude easy items that were solved by ≥ 90% of the participants (a product of the well-known ceiling-effects in Raven's SPM; Raven, 2000). This was a necessary step without whom factor analytic calculations would have been either not interpretable due to the lack of variance, or not feasible at all (see Brown, 2015). Consequently, 30 items (i.e., half of the tests items) had to be excluded from further analysis (the means of all excluded items are being provided in the Appendix). Multi-group confirmatory factor analyses (MGCFA) revealed invariant measurements of Raven's SPM over all cohorts as CFI changes did not exceed the threshold of 0.01 between the configural and the strict model (measurement invariance fit statistics are provided in Table 5).

Table 4

Pairwise comparisons of cohorts

Model	*df*	*F*	*p*	η^2
2001 vs. 2011	1	0.78	.377	0.001
2001 vs. 2018	1	6.80	.009	0.01
2001 vs. 2024	1	55.13	< .001	0.09
2011 vs. 2018	1	17.69	< .001	0.03
2011 vs. 2024	1	107.20	< .001	0.15
2018 vs. 2024	1	34.49	< .001	0.05

Note. df = Degrees of freedom

Table 5

Fit statistics for all cohorts and measurement invariance levels

Model	χ^2	*p*	*df*	CFI
Overall	871.832	< .001	405	0.979
2001	490.990	.002	405	0.987
2011	537.089	< .001	405	0.968
2018	480.459	.006	405	0.987
2025	462.332	.026	405	0.981
Configural	2070.76	< .001	1704	0.981
Strict	2232.377	< .001	1794	0.978

Note. df = Degrees of freedom; CFI = comparative fit index.

Standardized SPM score changes, measured in raw scores as well in latent means showed consistent increases from 2011 to 2024 after a stagnation from 2001 to 2011. Considering the cohorts that showed significant performance changes, raw score Cohen *d*'s were ranging from d = 0.21 to d = 0.84 and latent mean score Cohen *d*'s were situated in the range between d = 0.19 and d = 0.85, indicating small to large effects depending on the respective thresholds (see Table 6 for a detailed description). The trajectories of raw scores as well as latent means over the observed time period is depicted in Figure 1.

Table 6

Raw score- and latent mean-based effect sizes in Cohen d and DIQ-values

Cohort	2001	2011	2018	2024
2001	-	-0.14 (-2.10)	0.19 (1.68)*	0.61 (3.98)***
2011	-0.07 (-1.05)	-	0.36 (7.71)***	0.85 (9.81)***
2018	0.21 (1.85)***	0.33 (7.07)***	-	0.44 (11.00)***
2024	0.63 (4.11)***	0.84 (9.69)***	0.47 (11.75)***	-

Note. The bottom left triangular matric represents changes in the raw scores, the top right triangular matrix changes in latent means, values in parentheses are estimated DIQ values (IQ changes per decade). *p < .05; *** p < .001

Analyses of the incremental changes between measurement points revealed non-linear changes in test performance. Between 2001 and 2011, marginal, non-significant decreases in SPM scores were evident with losses of DIQ = -1.05 and DIQ = -2.10 for raw scores and latent means, respectively. Then, from 2011 to 2018, significant increases amounting to DIQ = 7.07 for raw scores and DIQ = 7.71 for latent means could be observed. Subsequently, performance gains grew even larger between 2018 and 2024, amounting to DIQ = 11.00 and DIQ = 11.75 for raw scores and latent means, respectively. Over the total time span, our data yielded increases of DIQ = 4.11 for raw scores, and DIQ = 3.98 for latent means.

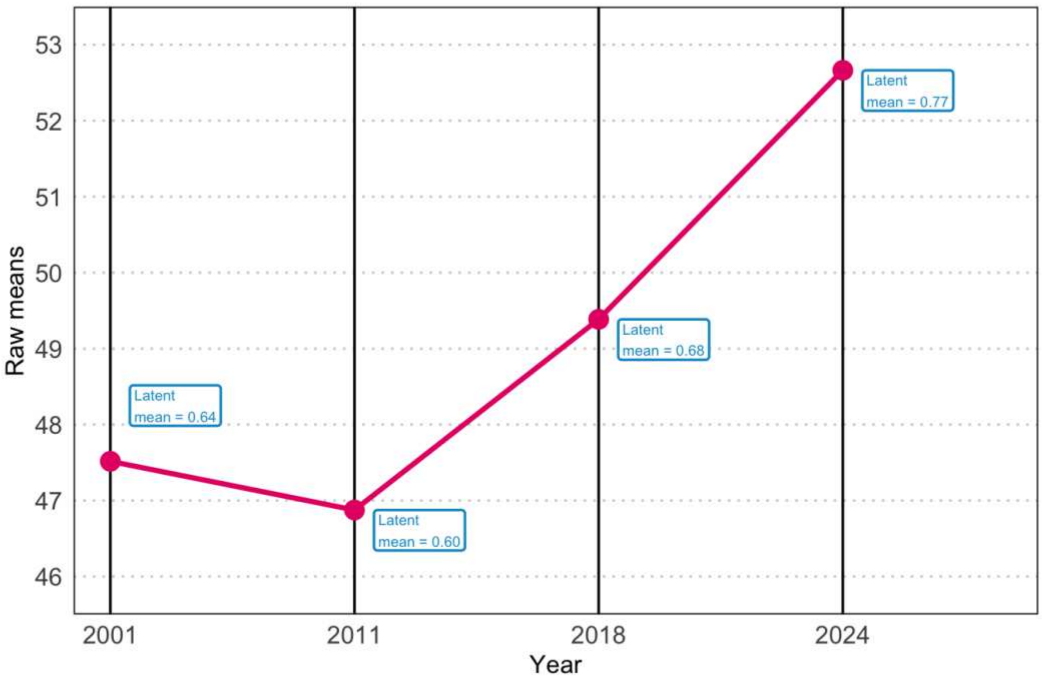


Figure 3. Raw (red) and latent (blue) mean score changes across the four cohorts

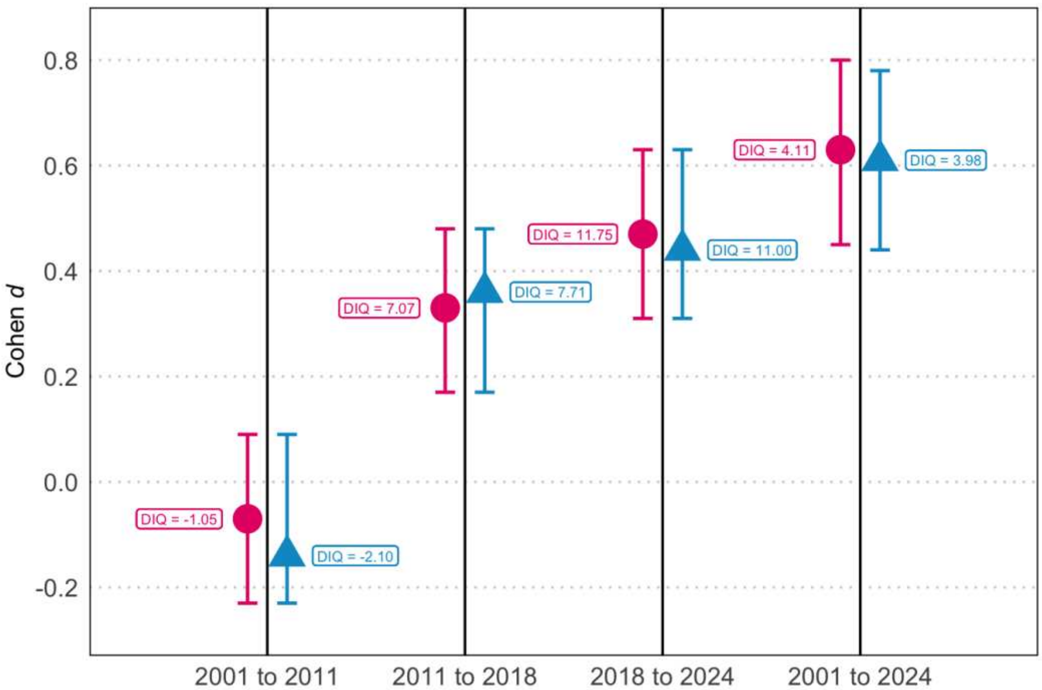


Figure 4. Cohen d and DIQ changes of raw (red) and latent (blue) scores between data collection years with 95% confidence intervals.

Discussion

We investigated cross-temporal changes in a measurement-invariant inductive reasoning task in Germanophone population-representative samples. Our analysis revealed a nonlinear, positive Flynn effect between 2001 and 2025 with measurement invariant changes amounting to 4.0 IQ points per decade (DIQ). At first glance, these findings fit in well within the existing body of literature regarding the domain, direction and extent of the effect (see Pietschnig & Voracek, 2015). The establishment of strict measurement invariance, however, stands out in contrast to the majority of investigations on the Flynn effect. The implications of our findings are discussed below.

Non-linear IQ trajectories

We could detect a slight (though non-significant) decrease in SPM scores between 2001 and 2011 amounting to DIQ = -2.1. Considering the widespread notion that Raven's SPM is one of the most influential instruments in defining the empirical status of positive test score changes (e.g., Nordmo et al., 2025), our finding of declining IQ scores in SPM seems rather unusual. It also contrasts findings from global meta-analyses revealing moderate performance increases for Raven's SPM (Brouwers et al., 2009).

Negative Flynn effects for Raven's SPM – despite being unusual – have been observed in several countries, such as Sudan (Batterjee & Ashira, 2015), Kuwait (Dutton et al., 2017) and Estonia (Kõrgesaar, 2013). These changes, however, could be partly attributed to country-specific properties (e.g., increased emphasis of religious content in Kuwaiti school curricula or overall poor education quality in Sudan) and can, therefore, not be generalized to other geographical contexts. Evidence from Germanophone standardization samples for Adaptive Matrices (AMT; Hornke et al., 1999) showed moderate test score decreases (DIQ = -2.3) between 2002 and 2005 (Lazaridis et al., 2022). At the same time, this very study also reported large test score increases in the same domain (G_f), highlighting different fluctuation rates between different subdomains on stratum I level (Lazaridis et al., 2022).

The slight IQ decline between 2001 and 2011 was subsequently accompanied by massive gains amounting to 7.7 IQ points per decade from 2011 to 2018 and 11.0

IQ points per decade between 2018 and 2024. Even though gains in matrices tests like Raven's SPM have historically been the highest in all of intelligence research (Flynn, 2012; Nordmo et al., 2025), the observed magnitudes are noteworthy, though not unsupported by the literature. Our findings align well with meta-analytical results detecting rising SPM scores amounting to DIQ = 7.0 (Trahan et al., 2014). With gain rates in Raven's SPM being about double the rate of other intelligence tests, rising SPM scores of such magnitude are not representative for a global Flynn effect numerically (see Trahan et al., 2014). Rather, by pointing to high test score volatility and the resulting structural instability in low *g*-tasks, the authors highlight the importance of differentiating between domain-specific and global IQ trajectories. Moreover, our pattern of non-meaningful changes followed by performance increases resonates well with recent evidence reporting stagnating SPM scores from 2001 to 2010 followed by performance gains from 2010 to 2017 (Lazaridis et al., 2022).

Another implication that can be drawn from this pattern is the differences in growth rates across assessment years. Even though performance declines, followed by -increases in G_f contradicts the idea of increasingly weaker gains in more recent years, they do underscore the non-linear nature of the Flynn effect (see Pietschnig & Voracek, 2015).

Taken together, our findings support the previously reported sensitivity of Raven's SPM to generational performance shifts. Furthermore, the observed changes in a fluid intelligence task of DIQ = 4.0 between 2001 and 2024 generally – even though variable across different IQ tests – align with earlier findings from the same domain (see Pietschnig & Voracek, 2015), supporting the notion of a positive Flynn effect.

(Sub)domain specificity

Even though our findings do resonate with the status quo of a positive Flynn effect in fluid intelligence, contrasting findings from the very same intelligence domain (G_f; see Dworak et al., 2023; Lazaridis et al., 2022; Oberleiter et al., 2024b) underscore the importance of not interpreting the present results as being consistent across the whole fluid ability domain. Rather, changes in different subdomains of G_f may take place independently, indicating that Flynn effect trends may be more complex than previously assumed.

Conceptualizing Flynn effect patterns in rather crude taxonomies like the G_f – G_c model (Cattell, 1963) or even considering them solely on a full-scale level (e.g., Trahan et al., 2014) might have masked more nuanced developments of subdomain-specific abilities as referred to in the CHC-model. It is, therefore, insufficient to conclude that performance changes in cognitive subdomains on stratum I level will change overall performance in the domain itself (fluid intelligence) on stratum II level. Consequently, it is crucial for future research to not understand Flynn effects as a universal indicator for increased or decreased intelligence, but rather in regard to more specific ability components embedded in more refined intelligence models.

Ability differentiation

Domain specific differences between contrasting findings within fluid intelligence (see Oberleiter et al., 2024b) suggest that cognitive abilities do not develop equally across different cohorts but may follow distinct trajectories. Specifically, it is possible that the observed ability changes in this study underly mechanisms of *ability differentiation* (Pietschnig et al., 2023). In line with this idea, societal developments lead to shifting environmental demands which, in turn, lead to the formation of different cohort dependent ability profiles. Thereby, different subdomains of fluid intelligence, like inductive reasoning (Raven's SPM used in this study) and e.g., figural reasoning (BEFKI; Wilhelm et al., 2014), as employed in a recent study (Oberleiter et al., 2024b), may be subject to differing environmental influences which distinctly impact their relevance and, thus, practice opportunities, i.e., the extent to which individuals are subjected to environmental demands and perform action in order to adapt to them. Consequently, figural reasoning might become increasingly more obsolete, as, for instance, solving a puzzle or navigating with a map have either ceased from everyday life overall or are adopted by modern digital applications (e.g., google maps), respectively (as theorized in Oberleiter et al., 2024b). Observations from Spanish students from the 2020s performing worse in visuospatial relations than those from the 1990s further support this claim (Colom et al., 2023).

Taken together, considerations on domain-specific trends of ability differentiation hint at the possibility that our findings do not indicate uniform gains across the entire G_f-domain or even general intelligence (*g*). Raven's SPM primarily captures inductive reasoning which is subordinated on stratum I level to fluid intelligence on stratum II level (Schneider & McGrew, 2018). In light of the broader body of Flynn literature with its inconsistent IQ trajectories across IQ domains in addition to recent findings that explicitly show ambiguous IQ trajectories within the same stratum II domain (see Lazaridis et al., 2022), it seems plausible that the observed performance gains in inductive reasoning do not extend across the entire domain of fluid intelligence. Moreover, the establishment of strict measurement invariance provides evidence for genuine ability changes instead of measurement artifacts. This does not mean, however, that the observed gains transcend to higher order factors, such as *g*. Rather, they underscore the importance of explicit investigation of which (sub)domain of intelligence is changing over time.

Genuine ability changes

We could establish strict measurement invariance across all assessment points, indicating equal factorial structure and -loadings, as well as equal residual invariance. As Item properties did not change over time, we can assume that the observed test score changes reflect genuine ability changes instead of measurement artifacts (such as DIF), thereby concluding that Raven’s SPM was indeed a valid measure for fluid intelligence in regard to inductive reasoning.

Our finding stands out methodologically against a wide array of studies investigating the Flynn effect using Raven’s SPM and comparable measures of fluid intelligence. Reported gains often stem from the mere comparison of raw scores without taking possible alterations in test taking behavior or DIF into consideration. This approach has fatal consequences, as previously reported Flynn effects may have been biased in their magnitude or even direction. Compared to our observed gains from 2001 to 2024 of DIQ = 4.0 – which align perfectly with meta-analytically proposed gain rates in fluid intelligence (DIQ = 4.1; e.g., Pietschnig & Voracek, 2013) – some previously reported SPM gains that did not account for measurement invariance deviate noticeably from this benchmark. For example, English school children have gained about 13.7 IQ points in 70 years for a rate of DIQ = 1.9 (Flynn, 2009). Interestingly, after 1979 gains have been declining with some age groups even suffering IQ losses. More recent data from Portuguese schoolkids using Raven’s

Colored Progressive Matrices (CPM; Raven et al., 1998) yield similar results (Carvalho et al., 2020). Gains were greater between 1991 and 2008 (DIQ = 2.6) than between 2008 and 2016 (DIQ = 1.7) including some age groups undergoing IQ losses. Considering the well documented non-linearity of the Flynn effect (e.g., Pietschnig & Voracek, 2013) a deceleration of gains in recent years is not as interesting as their low magnitude. Gains in range of 2 IQ points per decade would usually be expected in G_c domains, not in G_f domains (see Pietschnig & Voracek 2013). Other findings report substantially larger SPM gains. In Spain, IQ gains amounted to DIQ = 6.9 (Colom et al., 2005) with similar SPM gains being present in Argentina (Flynn & Rossi-Casé, 2012; DIQ = 6.1), where the authors even explicitly stated that due to unavailable data they ‘could not determine whether the gains were factor invariant’ (p. 146). An even larger Flynn effect was documented in China, where IQ increases amounted to DIQ = 8.3 (Wang & Lynn, 2018).

It is possible that, partly, previous studies reporting SPM gains are biased due to the lack of ensuring the temporal invariance of their measurement tool. On a full-scale IQ level, an example of the possible effects can be illustrated in regard to the documented absence of measurement invariance of over half of the items of different Wechsler scales (Gonthier & Grégoire, 2022). The bias inflicted by the DIF-items accumulated to an underestimation of the Flynn effect of DIQ = 3.0. In other words: Due to Wechsler items becoming more difficult over time the actual rate of IQ gains was higher by 3 IQ points per decade compared to the rate that did not consider item drift. Other fluid IQ measurements demonstrate the opposite pattern, meaning that the lack of measurement invariance proved to result in an overestimation of the Flynn effect, as some items got easier over time (Wicherts et al., 2004).

It remains unclear how many former studies on the Flynn effect are impacted by this issue. Still, evidence of DIF in prominent IQ tests indicates that biases of that sort are more than just hypothetical and plausibly affect a notable number of previously reported Flynn effects in both directions; some upward, some downward. These methodological considerations apply to Raven’s SPM as well, as it has been rarely subjected to explicit testing of measurement invariance.

These concerns are substantiated by the fact that many of the SPM-studies that actually conducted temporal measurement invariance analyses struggled to establish it appropriately. Younger generations tended to use differing problem-solving strategies for matrix reasoning than older generations (Fox & Mitchum, 2013). As a result, the evident SPM score differences between cohorts reflect qualitative differences in their solving strategies, not quantitative differences in the underlying inductive reasoning domain and are, therefore, not measurement invariant. Accordingly, it has been hypothesized that Raven's SPM is particularly susceptible to rule exposure (implicit exposure of solving strategies through education and cultural practices) and rule sensitivity (cognitive adaptation to the more complex environments), meaning that performance gains are rather a product of increased familiarity with abstract reasoning rules and greater flexibility in applying them than genuine gains in the underlying ability (Armstrong & Woodley, 2014). More recently, data from Norwegian armed forces suggests that for the most part of the 20th century IQ gains are mostly driven by increasing matrices test scores, with most of them not being measurement invariant (Nordmo et al., 2025). The authors conclude that 'temporal variations in intelligence scores can be more accurately attributed to shifts in test properties […]' (p. 7) than true ability changes. Other measurement tools also struggle to establish measurement invariance, for example the NLSY79 (Beaujean & Osterlind, 2009) or the Estonian National Intelligence test (Must et al., 2009).

In light of these methodological challenges, our study stands out as an important counterexample to the majority of studies failing to ascribe measurement invariant test properties to Raven's SPM. Unlike many prior investigations, we could show that our observed score gains are not driven by increased item familiarity, differing test taking strategies or shifting item properties, but reflect real performance gains in inductive reasoning. Based on our findings we can conclude that Raven's SPM – at least in Austria – can be a valid measurement tool for detecting genuine performance gains in inductive reasoning and fluid intelligence. This interpretation resonates well with recent evidence from over 12,000 Austrian residents exhibiting strict measurement invariant SPM score changes (Lazaridis et al., 2022).

We cannot, however, definitively assume that this is the case for other geographical regions, as temporal measurement invariance does not automatically imply geographical measurement invariance. Evidence from sub-Saharan African countries illustrates this point by raising questions about transnational comparability of SPM scores and their culture fairness (Wicherts et al., 2010). Raven's SPM did not display the same factorial structure as in western samples, implying that different constructs were being measured depending on geographical context. Factor analyses and IRT-modelling showed that the large gap between African (IQ = 80) and US

(IQ = 100) norms were attributable to test validity- and measurement invariance issues and not to differences in general intelligence. Besides the fact that Raven's SPM – contrary to what is stated in the manual (see Raven et al., 2018) – does not seem to capture intelligence unidimensionally (e.g., Lynn et al., 2004; van der Ven & Ellis, 2000), it also fails to measure invariantly across genders (Mackintosh & Bennett, 2005). Accordingly, it has been shown that different the Wechsler scales (e.g., WAIS, WISC; Wechsler, 1974, 1981) do not measure intelligence across different subpopulations invariantly (e.g., Winter et al., 2024; Beaujean & Sheng, 2014).

Considerations of that sort underscore the importance of proper methodology when conducting Flynn effect research. When assessing IQ changes it is essential to determine, whether the measurement of choice measured the specific IQ domain invariantly over time. Without such analyses we cannot draw valid comparisons across cohorts and measurement points. If not given, differences in IQ trends may rather reflect methodological omissions by the researchers instead of actual ability changes. A useful analogy by Jensen (1994; see also Gonthier & Grégoire, 2022) helps to illustrate this issue: Inferring changes of intelligence based on changes in test scores is equivalent to inferring differences in height based on differences in the length of people's shadows. Comparing the length at one particular point in time will yield correct results, but comparing the shadow's length in different seasons, when the sun's position on the horizon is either lower or higher, may yield biased estimates of height differences. In similar manner, cross-cultural comparisons of SPM scores should undergo explicit invariance testing, as differences in test scores may be biased by culturally or structurally determined DIF.

Limitations

It has to be noted that, partly, the observed temporal fluctuations of SPM scores (e.g., a non-significant decline between 2001 and 2011 followed by large gains until 2024) may reflect systematic group differences in age, sex and level of education between almost all cohorts – except the purposefully collected cohort of 2024 and its 2001 stratification template (detailed statistics are provided in the Appendix). The 2011 cohort's mean level of education is significantly lower compared to all other cohorts, possibly accounting for the slight decrease prior and the subsequent increases in later cohorts. Additionally, the mean age of the 2018 cohort is by far the highest. Even though age moderating effects, remain, as previously mentioned, debated, a claim could be made that older participants represented in the 2018 cohort may have attenuated the observed SPM scores in comparison to the younger 2024 cohort (see Wongupparaj et al., 2023).Therefore, a valid conclusion about the extent of the Flynn effect-related changes in ability can only be drawn on the basis of a comparison between the 2001 and 2024 cohorts, which do not differ significantly regarding sociodemographics.

Conclusion

Our analysis compared three archival cohorts from 2001, 2011 and 2018 with a purposefully collected sample from 2024 representative in regard to the demographic composition of year 2001. We used Raven's SPM (Raven et al., 2018) as our primary measurement of fluid intelligence in order to detect meaningful IQ score changes across the four cohorts.

Over the total observed timespan, we detected a Flynn effect of DIQ = 4.0. While the margin of IQ gains fits in well with previously reported Flynn effects in fluid intelligence (e.g., Pietschnig & Voracek, 2015), the conclusion of a universally positive Flynn effect in that domain may be flawed. Mechanisms of ability differentiation suggest that the observed gains may not be generalizable to the whole domain of fluid intelligence but are taking place domain-specifically in inductive reasoning on stratum I level (see Pietschnig et al., 2023).

As strict measurement invariance could be established across all cohorts, the observed changes reflect genuine ability changes instead of measurement artifacts, such as DIF. Raven's SPM proved to be a viable measurement tool for detecting a positive Flynn effect, providing robust evidence for increases in inductive reasoning over the past two decades in Austria. In order to interpret the Flynn effect properly, it is crucial for future research to ensure temporal invariance of the measurement tool.

References

Al-Shahomee, A. A., Abdalla, S. E.-G., & Lynn, R. (2018). An increase of intelligence in Libya from 2008 to 2017. *Personality and Individual Differences*, *130*, 147–149. https://doi.org/10.1016/j.paid.2018.04.004

Andrzejewski, D., Oberleiter, S., Vetter, M., & Pietschnig, J. (2024). Increasing IQ Test Scores and Decreasing g: The Flynn Effect and Decreasing Positive Manifold Strengths in Austria (2005-2018). *Journal of Intelligence*, *12*(12), 130. https://doi.org/10.3390/jintelligence12120130

Armstrong, E. L., & Woodley, M. A. (2014). The rule-dependence model explains the commonalities between the Flynn effect and IQ gains via retesting. *Learning and Individual Differences*, *29*, 41–49. https://doi.org/10.1016/j.lindif.2013.10.009

Bates, T. (2013, 9th February). *Carroll three stratum model of human Intelligence* [PNG]. Retrieved from https://commons.wikimedia.org/wiki/File:Carroll_three_stratum_model_of_human_Intelligence.png

Batterjee, A. A. & Ashria, I. (2015). Intelligence and Education: The Sudan Case. *Mankind Quarterly*, *56*(2), 136-167. https://doi.org/10.46469/mq.2015.56.2.3

Beaujean, A. A., & Osterlind, S. J. (2008). Using item response theory to assess the Flynn effect in the National Longitudinal Study of youth 79 children and young adults data. *Intelligence, 36*(5), 455–463. https://doi.org/10.1016/j.intell.2007.10.004

Beaujean, A., & Sheng, Y. (2014). Assessing the Flynn Effect in the Wechsler Scales. *Journal of Individual Differences*, *35*(2), 63–78. https://doi.org/10.1027/1614-0001/a000128

Blum, F., Didi, H.-J., Fay, E., Maichle, U., Trost, G., Wahlen, H.-J., & Gittler, G. (2018). *Intelligenz-Basis-Funktionen (IBF) [Basic Intelligence Functions (IBF)].* Schuhfried.

Bradmetz, J., & Mathy, F. (2006). An estimate of the Flynn Effect: changes in IQ and subtest gains of 10-yr-old French children between 1965 and 1988. *Psychological reports*, *99*(3), 743–746. https://doi.org/10.2466/PR0.99.3.743-746

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

Breit, M., Scherrer, V., Tucker-Drob, E. M., & Preckel, F. (2024). The stability of cognitive abilities: A meta-analytic review of longitudinal studies. *Psychological bulletin*, *150*(4), 399–439. https://doi.org/10.1037/bul0000425

Brouwers, S. A., Van de Vijver, F. J. R., & Van Hemert, D. A. (2009). Variation in Raven's Progressive Matrices scores across time and place. *Learning and Individual Differences*, *19*(3), 330–338. https://doi.org/10.1016/j.lindif.2008.10.006

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

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

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

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

Colom, R., Andrés-Pueyo, A., & Juan-Espinosa, M. (1998). Generational IQ gains: Spanish data. *Personality and Individual Differences*, *25*(5), 927–935. https://doi.org/10.1016/S0191-8869(98)00090-7

Colom, R., Flores-Mendoza, C. E., & Abad, F. J. (2007). Generational changes on the draw-a-man test: A comparison of Brazilian urban and rural children tested in 1930, 2002 and 2004. *Journal of Biosocial Science*, *39*(1), 79–89. https://doi.org/10.1017/S0021932005001173

Colom, R., García, L. F., Shih, P. C., & Abad, F. J. (2023). Generational intelligence tests score changes in Spain: Are we asking the right question? *Intelligence*, *99*, 101772. https://doi.org/10.1016/j.intell.2023.101772

Colom, R., Lluis-Font, J. M. a, & Andrés-Pueyo, A. (2005). The generational intelligence gains are caused by decreasing variance in the lower half of the distribution: Supporting evidence for the nutrition hypothesis. *Intelligence*, *33*(1), 83–91. https://doi.org/10.1016/j.intell.2004.07.010

Coolberth, Nicole L. 2024. *Digital Equity Advances in Congress: Digital Skills for Today's Workforce*. National Skills Coalition. https://nationalskillscoalition.org/blog/future-of-work/digital-equity-advances-in-congress-digital-skills-for-todays- workforce/ (accessed on 28 June 2025).

Deoni, S. (2022). Impact of the COVID-19 Pandemic Environment on Early Child Brain and Cognitive Development. *Biological Psychiatry (1969)*, *91*(9), 26. https://doi.org/10.1016/j.biopsych.2022.02.082

Dickens, W. T., & Flynn, J. R. (2001). Heritability Estimates Versus Large Environmental Effects: The IQ Paradox Resolved. *Psychological Review*, *108*(2), 346–369. https://doi.org/10.1037/0033-295X.108.2.346

Dutton, E., Bakhiet, S. F., Essa, Y. A. S., Blahmar, T. A. M., & Hakami, S. M. A. (2017). A Negative Flynn Effect in Kuwait: The same effect as in Europe but with seemingly different causes. *Personality and Individual Differences*, *114*, 69–72. https://doi.org/10.1016/j.paid.2017.03.060

Dutton, E., Bakhiet, S. F. A., Osman, H. A., Becker, D., Essa, Y. A. S., Blahmar, T. A. M., Lynn, R., & Hakami, S. M. (2018). A Flynn Effect in Khartoum, the Sudanese capital, 2004–2016. *Intelligence*, *68*, 82–86. https://doi.org/10.1016/j.intell.2018.03.007

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

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

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

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

Elbanna, L. M. A., Bakhiet, S. F. A., Ali, S. A. T., Cheng, H., & Lynn, R. (2018). Gender differences in intelligence in a sample of 5-12 year olds in Sudan. *Mankind Quarterly, 59*, 87–91.

Emanuelsson, I., Reuterberg, S.-E., & Svensson, A. (1993). Changing Differences in Intelligence? Comparisons between groups of 13-year-olds tested from 1960 to 1990. *Scandinavian Journal of Educational Research*, *37*(4), 259–277. https://doi.org/10.1080/0031383930370401

Eppig, C., Fincher, C. L., & Thornhill, R. (2010). Parasite prevalence and the worldwide distribution of cognitive ability. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1701), 3801–3808. https://doi.org/10.1098/rspb.2010.0973

European Commission. 2020. *Commission Presents European Skills Agenda* [Text]. European Commission–European Commission. https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1196 (accessed on 01 July 2025).

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., Frans, Ö., Gintiliené, G., Hagemeister, C., Halama, P., Iliescu, D., Jaworowska, A., Jiménez, P., Manthouli, M., Matesic, K., Schittekatte, M., Sümer, H. C., & Urbánek, T. (2012). Testing practices in the 21st century: Developments and European psychologists' opinions. *European Psychologist, 17*(4), 300–319. https://doi.org/10.1027/1016-9040/a000102

Eysenck, H. J. (1988). The concept of "intelligence": Useful or useless? *Intelligence*, *12*(1), 1–16. https://doi.org/10.1016/0160-2896(88)90019-0

Flieller, A. (1988). Application du modèle de Rasch à un problème de comparaison de générations. *Bulletin de psychologie*, *42*(388), 86–91. https://doi.org/10.3406/bupsy.1988.12973

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

Flynn, J. R. (1987). Massive IQ Gains in 14 Nations: What IQ Tests Really Measure. *Psychological Bulletin*, *101*(2), 171–191. https://doi.org/10.1037/0033-2909.101.2.171

Flynn, J. R. (1998). IQ gains over time: Toward finding the causes. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 25–66). American Psychological Association. https://doi.org/10.1037/10270-001

Flynn, J. R. (1999). Searching for Justice: The Discovery of IQ Gains Over Time. *The American Psychologist*, *54*(1), 5–20. https://doi.org/10.1037/0003-066X.54.1.5

Flynn, J. R. (2007). *What is intelligence? Beyond the Flynn effect.* Cambridge University Press. https://doi.org/10.1017/CBO9780511605253

Flynn J. R. (2009). Requiem for nutrition as the cause of IQ gains: Raven's gains in Britain 1938-2008. *Economics and human biology*, *7*(1), 18–27. https://doi.org/10.1016/j.ehb.2009.01.009

Flynn, J. R. (2012). *Are We Getting Smarter?: Rising IQ in the Twenty-First Century*. Cambridge University Press. https://doi.org/10.1017/CBO9781139235679

Flynn, J. R. (2013). The "Flynn Effect" and Flynn's paradox. *Intelligence, 41*(6), 851–857. https://doi.org/10.1016/j.intell.2013.06.014

Flynn, J. R., & Shayer, M. (2018). IQ decline and Piaget: Does the rot start at the top? *Intelligence*, *66*, 112–121. https://doi.org/10.1016/j.intell.2017.11.010

Fox, M. C., & Mitchum, A. L. (2013). A Knowledge-Based Theory of Rising Scores on "Culture-Free" Tests. *Journal of Experimental Psychology. General*, *142*(3), 979–1000. https://doi.org/10.1037/a0030155

Gardner, H. (1993). *Frames of mind: the theory of multiple intelligences* (2. ed. with new introd.). Fontana Press.

Gignac, G. E. (2015). Raven's is not a pure measure of general intelligence: Implications for g factor theory and the brief measurement of g. *Intelligence*, *52*, 71–79. https://doi.org/10.1016/j.intell.2015.07.006

Gittler, G. (1984). Entwicklung und Erprobung eines neuen Testinstruments zur Messung des räumlichen Vorstellungsvermögens [Development and test of a new test mea- sure of spatial task performance]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 5*, 141–165.

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

Gonthier, C., Grégoire, J., & Besançon, M. (2021). No negative Flynn effect in France: Why variations of intelligence should not be assessed using tests based on cultural knowledge. *Intelligence*, *84*, 101512. https://doi.org/10.1016/j.intell.2020.101512

Gorey, K. M. (2001). Early Childhood Education: A Meta-Analytic Affirmation of the Short- and Long-Term Benefits of Educational Opportunity. *School Psychology Quarterly*, *16*(1), 9–30. https://doi.org/10.1521/scpq.16.1.9.19163

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, *24*(1), 13–23. https://doi.org/10.1016/S0160-2896(97)90011-8

Gottfredson, L., & Saklofske, D. H. (2009). Intelligence: Foundations and Issues in Assessment. *Canadian Psychology = Psychologie Canadienne*, *50*(3), 183–195. https://doi.org/10.1037/a0016641

Gould, S. J. (1981). *The mismeasure of man*. Norton.

Gunnesch-Luca, G., & Iliescu, D. (2020). Time and generational changes in cognitive performance in Romania. *Intelligence*, *79*, 101430–101439. https://doi.org/10.1016/j.intell.2020.101430

Haier, R. J., Siegel, B. V., Nuechterlein, K. H., Hazlett, E., Wu, J. C., Paek, J., Browning, H. L., & Buchsbaum, M. S. (1988). Cortical glucose metabolic rate correlates of abstract reasoning and attention studied with positron emission tomography. *Intelligence*, *12*(2), 199–217. https://doi.org/10.1016/0160-2896(88)90016-5

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*(5), 253–270. https://doi.org/10.1037/h0023816

Hõrak, P. & Valge, M. (2015). Why did children grow so well at hard times? The ultimate importance of pathogen control during puberty. *Evolution Medicine And Public Health*, *2015*(1), 167–178. https://doi.org/10.1093/emph/eov017

Hornke, L. F., Etzel, S., & Rettig, K. (1999). *AMT Adaptive Matrices Test*. APA PsycTests. https://doi.org/10.1037/t10889-000

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures*. Basic Books.

Jensen, A. R. (1999). The G factor: The science of mental ability. *Psycholoqy*, 10.

Jones, H. E., & Conrad, H. S. (1933). The growth and decline of intelligence: a study of a homogeneous group between the ages of ten and sixty. *Genetic Psychology Monographs*.

Kanaya, T., Scullin, M. H., & Ceci, S. J. (2003). The Flynn Effect and U.S. Policies: The Impact of Rising IQ Scores on American Society Via Mental Retardation Diagnoses. *The American Psychologist*, *58*(10), 778–790. https://doi.org/10.1037/0003-066X.58.10.778

Koivunen, S. (2007). *Suomalaismiesten kognitiivisen kykyprofiilin muutokset 1988-2001: Flynnin efektiä suomalaisessa aineistossa*? (Pro Gradu) Jyväskylä University.

Kõrgesaar, M. (2013). *Flynni Efekti Esinemine Eesti Abiturientide Seas Raveni Testi Pohjal*. Tartu Ulikool. http://hdl.handle.net/10062/30644

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

Lucas, A., Morley, R., & Cole, T. J. (1998). Randomised trial of early diet in preterm babies and later intelligence quotient. *BMJ (Clinical research ed.)*, *317*(7171), 1481–1487. https://doi.org/10.1136/bmj.317.7171.1481

Lynn, R. (1990). Differential rates of secular increase of five major primary abilities. *Biodemography and Social Biology*, *37*(1–2), 137–141. https://doi.org/10.1080/19485565.1990.9988753

Lynn, R. (2006). *Race differences in intelligence: An evolutionary analysis.* Washington Summit Publishers.

Lynn, R. (2009). Fluid intelligence but not vocabulary has increased in Britain, 1979–2008. *Intelligence*, *37*(3), 249–255. https://doi.org/10.1016/j.intell.2008.09.007

Lynn, R., Allik, J., & Irwing, P. (2004). Sex differences on three factors identified in Raven's Standard Progressive Matrices. *Intelligence*, *32*(4), 411–424. https://doi.org/10.1016/j.intell.2004.06.007

Lynn, R., & Cooper, C. (1994). A secular decline in the strength of Spearman's g in Japan. *Current Psychology*, *13*(1), 3–9. https://doi.org/10.1007/BF02686854

Lynn, R., & Hampson, S. (1986). The rise of national intelligence: Evidence from Britain, Japan and the U.S.A. *Personality and Individual Differences*, *7*(1), 23–32. https://doi.org/10.1016/0191-8869(86)90104-2

Lynn, R., Hampson, S. L. & Mullineux, J. C. (1987). A long-term increase in the fluid intelligence of English children. *Nature*, *328*(6133), 797. https://doi.org/10.1038/328797a0

Mackintosh, N. & Bennett, E. (2005). What do Raven's Matrices measure? An analysis in terms of sex differences. *Intelligence*, *33*(6), 663–674. https://doi.org/10.1016/j.intell.2005.03.004

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative Longitudinal Structural Analyses of the Growth and Decline of Multiple Intellectual Abilities Over the Life Span. *Developmental Psychology*, *38*(1), 115–142. https://doi.org/10.1037/0012-1649.38.1.115

Meisenberg, G. (2003). IQ Population Genetics: It's not as Simple as You Think. *Mankind Quarterly, 44*(2), 185–210. https://doi.org/10.46469/mq.2003.44.2.4

Meisenberg, G., & Lynn, R. (2023). Ongoing trends of human intelligence. *Intelligence*, *96*, 101708. https://doi.org/10.1016/j.intell.2022.101708

Meredith, W. (1993). Measurement Invariance, Factor Analysis and Factorial Invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Merrill, M. A. (1938). The significance of IQ's on the revised Stanford-Binet scales. *Journal of Educational Psychology, 29*(9), 641–651. https://doi.org/10.1037/h0057523

Merten, N., Pinto, A. A., Paulsen, A. J., Chen, Y., Schubert, C. R., & Cruickshanks, K. J. (2022). Better cognitive function in younger generations - Insights from two cohort studies of middle-aged to older adults in Wisconsin. *Maturitas*, *162*, 31–36. https://doi.org/10.1016/j.maturitas.2022.04.002

Mingroni, M. A. (2007). Resolving the IQ Paradox: Heterosis as a Cause of the Flynn Effect and Other Trends. *Psychological Review*, *114*(3), 806–829. https://doi.org/10.1037/0033-295X.114.3.806

Moehring, A., Schroeders, U., Leichtmann, B., & Wilhelm, O. (2016). Ecological momentary assessment of digital literacy: Influence of fluid and crystallized intelligence, domain-specific knowledge, and computer usage. *Intelligence*, *59*, 170–180. https://doi.org/10.1016/j.intell.2016.10.003

Must, O., te Nijenhuis, J., Must, A., & van Vianen, A. E. M. (2009). Comparability of IQ scores over time. *Intelligence*, *37*(1), 25–33. https://doi.org/10.1016/j.intell.2008.05.002

Neisser, U. (1998). *The rising curve : long-term gains in IQ and related measures* (1. ed.). American Psychological Ass.

Nordmo, M., Norrøne, T. N., & Lang-Ree, O. C. (2025). Reevaluating the Flynn effect, and the reversal: Temporal trends and measurement invariance in Norwegian armed forces intelligence scores. *Intelligence*, *110*, 101909. https://doi.org/10.1016/j.intell.2025.101909

Oberleiter, S., Fries, J., Dejardin, F., Heller, J., Schaible, C., Vetter, M., Voracek, M., & Pietschnig, J. (2024a). Inconsistent Flynn effect patterns may be due to a decreasing positive manifold: Cohort-based measurement-invariant IQ test score changes from 2005 to 2024. *Intelligence*, *107*, 101867. https://doi.org/10.1016/j.intell.2024.101867

Oberleiter, S., Patzl, S., Fries, J., Diedrich, J., Voracek, M., & Pietschnig, J. (2024b). Measurement-Invariant Fluid Anti-Flynn Effects in Population—Representative German Student Samples (2012–2022). *Journal of Intelligence*, *12*(1), 9. https://doi.org/10.3390/jintelligence12010009

Oberleiter, S., Wurzer, J., Mikas, M., Held, M., Wieland, B., Zeilinger, E. L., Voracek, M., & Pietschnig, J. (2025). Generational IQ test score changes and the positive manifold of intelligence: evidence from Austrian Air Force pilots and air traffic controllers (1992–2016). *Frontiers in Psychology*, *16*, 1547520. https://doi.org/10.3389/fpsyg.2025.1547520

Parker, S. (2014). Were the Victorians cleverer than us? Maybe, maybe not. *Intelligence*, *47*, 1–2. https://doi.org/10.1016/j.intell.2014.08.002

Piaget, J. (1985). *The equilibration of cognitive structures: The central problem of intellectual development*. Chicago: University of Chicago Press.

Piaget, J., & Inhelder, B. (1974). *The child's construction of quantities: Conservation and atomism*. Routledge & Kegan Paul.

Pietschnig, J. (2016). The Flynn Effect: Technology May Be Part of It, But Is Most Certainly Not All of It. *Measurement Interdisciplinary Research And Perspectives*, *14*(2), 70–73. https://doi.org/10.1080/15366367.2016.1171612

Pietschnig, J., 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. https://doi.org/10.1016/j.intell.2021.101544

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. https://doi.org/10.1016/j.intell.2015.10.004

Pietschnig, J., Oberleiter, S., Toffalini, E., & Giofrè, D. (2023). Reliability of the g factor over time in Italian INVALSI data (2010-2022): What can achievement-g tell us about the Flynn effect? *Personality and Individual Differences*, *214*, 112345. https://doi.org/10.1016/j.paid.2023.112345

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

Pietschnig, J., & Voracek, M. (2015). One Century of Global IQ Gains. *Perspectives on Psychological Science*, *10*(3), 282–306. https://doi.org/10.1177/1745691615577701

Pietschnig, J., Voracek, M., & Formann, A. K. (2010). Pervasiveness of the IQ Rise: A Cross-Temporal Meta-Analysis. *PloS One*, *5*(12), e14406–e14406. https://doi.org/10.1371/journal.pone.0014406

Pietschnig, J., Voracek, M., and Gittler, G. (2018). Is the Flynn effect related to migration? Meta-analytic evidence for correlates of stagnation and reversal of generational IQ test score changes. *Politische Psychol. 6*, 276–283.

Platt, J. M., Keyes, K. M., McLaughlin, K. A. & Kaufman, A. S. (2019). The Flynn effect for fluid IQ may not generalize to all ages or ability levels: A population-based study of 10,000 US adolescents. *Intelligence*, *77*, 101385. https://doi.org/10.1016/j.intell.2019.101385

Qi, Y., & Xiong, Y. (2023). Intercohort upsurge of cognitive ability among the general population in China: Evaluating a Flynn effect. *Intelligence*, *98*, 101752. https://doi.org/10.1016/j.intell.2023.101752

Raven, J. C. (1938). *Progressive Matrices.* H. K. Lewis & Co., Ltd.

Raven, J. (2000). The Raven's Progressive Matrices: Change and Stability over Culture and Time. *Cognitive Psychology*, *41*(1), 1–48. https://doi.org/10.1006/cogp.1999.0735

Raven, J., Raven, J. C., & Court, J. H. (1998). Section 2: Coloured Progressive Matrices (1998 edition). Introducing the parallel version of the test. In *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Oxford Psychologists Press.

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

R Core Team (2025). *R: A Language and Environment for Statistical Computing.* Vienna: R Core Team. https://www.R-project.org/ (accessed on 22 July 2025).

Yves Rosseel (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1-36. URL http://www.jstatsoft.org/v48/i02/

Rindermann, H., Becker, D., & Coyle, T. R. (2017). Survey of expert opinion on intelligence: The FLynn effect and the future of intelligence. *Personality and Individual Differences*, *106*, 242–247. https://doi.org/10.1016/j.paid.2016.10.061

Rindermann, H., & Thompson, J. (2013). Ability rise in NAEP and narrowing ethnic gaps? *Intelligence*, *41*(6), 821–831. https://doi.org/10.1016/j.intell.2013.06.016

Rindermann, H., & Thompson, J. (2016). The cognitive competences of immigrant and native students across the world: An analysis of gaps, possible causes and impact. *Journal of Biosocial Science, 48,* 66–93. http://dx.doi.org/10.1017/S0021932014000480

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

Rodgers, J. L. (2023). Eleven articles and 27 authors pay tribute to James Flynn: A summary and critique of special issue articles on the Flynn effect. *Intelligence*, *101*, 101794. https://doi.org/10.1016/j.intell.2023.101794

Rodgers, J. L., Wänström, L., & Ang, S. (2025). Putting the Flynn effect under the microscope: Item-level patterns in NLSYC PIAT-math scores, 1986–2004. *Intelligence*, *109*, 101897. https://doi.org/10.1016/j.intell.2024.101897

Russell, E. W. (2007). The Flynn Effect Revisited. *Applied Neuropsychology*, *14*(4), 262–266. https://doi.org/10.1080/09084280701719211

Schaie, K. W. (1958). Rigidity-flexibility and intelligence: A cross-sectional study of the adult life span from 20 to 70 years. *Psychological Monographs: general and applied*, *72*(9), 1.

Schaie, K. W., & Strother, C. R. (1968). A cross-sequential study of age changes in cognitive behavior. *Psychological Bulletin*, *70*(6, Pt.1), 671–680. https://doi.org/10.1037/h0026811

Schneider, W. J., & McGrew, K. S. (2018). The Cattell–Horn–Carroll theory of cognitive abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73–163). The Guilford Press.

Shayer, M., Ginsburg, D., & Coe, R. (2007). Thirty years on - a large anti-Flynn effect? The Piagetian test Volume & Heaviness norms 1975-2003. *British Journal of Educational Psychology*, *77*(1), 25–41. https://doi.org/10.1348/000709906X96987

Shayer, M., & Ginsburg, D. (2009). Thirty years on - a large anti-Flynn effect/ (II): 13- and 14-year-olds. Piagetian tests of formal operations norms 1976-2006/7. *British Journal of Educational Psychology*, *79*(3), 409–418. https://doi.org/10.1348/978185408X383123

Suna, H. E., & Ozer, M. (2024). Medium- and long-term outcomes of early childhood education: experiences from Turkish large-scale assessments. *Humanities & Social Sciences Communications*, *11*(1), 1–15. https://doi.org/10.1057/s41599-024-03241-9

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

Sundet, J. M., Eriksen, W., Borren, I. & Tambs, K. (2010). The Flynn effect in sibships: Investigating the role of age differences between siblings. *Intelligence*, *38*(1), 38–44. https://doi.org/10.1016/j.intell.2009.11.005

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107

Teasdale, T. W., & Owen, D. R. (1987). National secular trends in intelligence and education: a twenty-year cross-sectional study. *Nature*, *325*(6100), 119–121. https://doi.org/10.1038/325119a0

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

te Nijenhuis, J., & van der Flier, H. (2013). Is the Flynn effect on g?: A meta-analysis. *Intelligence*, *41*(6), 802–807. https://doi.org/10.1016/j.intell.2013.03.001

Trahan, L. H., Stuebing, K. K., Fletcher, J. M., & Hiscock, M. (2014). The Flynn Effect: A Meta-Analysis. *Psychological Bulletin*, *140*(5), 1332–1360. https://doi.org/10.1037/a0037173

Tran, U. S., Hofer, A. A., & Voracek, M. (2014). Sex differences in general knowledge: Meta-analysis and new data on the contribution of school-related moderators among high-school students. *PLoS One, 9*(10), Article e110391. https://doi.org/10.1371/journal.pone.0110391

Tuddenham, R. D. (1948). Soldier intelligence in World Wars I and II. *The American Psychologist*, *3*(2), 54–56. https://doi.org/10.1037/h0054962

van der Ven, A. H. G. S., & Ellis, J. L. (2000). A Rasch analysis of Raven's standard progressive matrices. *Personality and Individual Differences*, *29*(1), 45–64. https://doi.org/10.1016/S0191-8869(99)00177-4

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Visser, B. A., Ashton, M. C., & Vernon, P. A. (2006a). Beyond g: Putting multiple intelligences theory to the test. *Intelligence*, *34*(5), 487–502. https://doi.org/10.1016/j.intell.2006.02.004

Visser, B. A., Ashton, M. C., & Vernon, P. A. (2006b). g and the measurement of Multiple Intelligences: A response to Gardner. *Intelligence*, *34*(5), 507–510. https://doi.org/10.1016/j.intell.2006.04.006

Wang, M., & Lynn, R. (2018). Intelligence in the People's Republic of China. *Personality and Individual Differences*, *134*, 275–277. https://doi.org/10.1016/j.paid.2018.06.010

Warrag, H., Khaleefa, O., Al-Thani, A., & Lynn, R. (2018). An increase of intelligence in Qatar from 1986 to 2011. *Mankind Quarterly, 59*, 115–120.

Wechsler, D. (1974). *Wechsler intelligence scale for children, revised : WISC-R ; manual*. Psychological Corp.

Wechsler, D. (1981). *WAIS-R : Wechsler adult intelligence scale - revised.* Psychological Corp.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale--Fourth Edition (WAIS-IV)* [Database record]. APA PsycTests. https://doi.org/10.1037/t15169-000

Wicherts, J. M. (2007). *Group differences in intelligence test performance [Unpublished dissertation]*. University of Amsterdam. Retrieved from https://pure.uva.nl/ws/files/4175964/46967_Wicherts.pdf

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*(5), 509–537. https://doi.org/10.1016/j.intell.2004.07.002

Wicherts, J. M., Dolan, C. V., Carlson, J. S., & van der Maas, H. L. J. (2010). Raven's test performance of sub-Saharan Africans: Average performance, psychometric properties, and the Flynn Effect. *Learning and Individual Differences*, *20*(3), 135–151. https://doi.org/10.1016/j.lindif.2009.12.001

Wilhelm, O., Schroeders, U., & Schipolowski, S. (2014). *BEFKI 8-10 : Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 8. bis 10. Jahrgangsstufe ; Manual*. Hogrefe.

Williams, R. L. (2013). Overview of the Flynn effect. *Intelligence*, *41*(6), 753–764. https://doi.org/10.1016/j.intell.2013.04.010

Winter, E. L., Trudel, S. M., & Kaufman, A. S. (2024). Wait, Where's the Flynn Effect on the WAIS-5? *Journal of Intelligence*, *12*(11), 118. https://doi.org/10.3390/jintelligence12110118

Wongupparaj, P., Wongupparaj, R., Morris, R. G., & Kumari, V. (2023). Seventy years, 1000 samples, and 300,000 SPM scores: A new meta-analysis of Flynn effect patterns. *Intelligence*, *98*, 101750. https://doi.org/10.1016/j.intell.2023.101750

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. https://doi.org/10.3389/fnhum.2015.00131

Woodley, M. A. (2011). Heterosis Doesn't Cause the Flynn Effect: A Critical Examination of Mingroni (2007). *Psychological Review*, *118*(4), 689–693. https://doi.org/10.1037/a0024759

Woodley, M. A. (2012). A life history model of the Lynn—Flynn effect: Special Issue on Evolution of race and sexe differences in intelligence and personality: Tribute to Richard Lynn at eighty. *Personality and Individual Differences*, *53*(2), 152–156.

Woodley, M. A., & Meisenberg, G. (2013). In the Netherlands the anti-Flynn effect is a Jensen effect. *Personality and Individual Differences*, *54*(8), 871–876. https://doi.org/10.1016/j.paid.2012.12.022

Woodley, M. A., Figueredo, A. J., Brown, S. D. & Ross, K. C. (2013). Four successful tests of the Cognitive Differentiation–Integration Effort hypothesis. *Intelligence*, *41*(6), 832–842. https://doi.org/10.1016/j.intell.2013.02.002

Woodley, M. A., Fieder, M., Sarraf, M. A., & Peñaherrera-Aguirre, M. (2024). Requiem for Heterosis as a Cause of the Flynn Effect: Positive Combined Effects of Numbers and Lengths of Homozygosity Runs on Offspring-Parent Differences in Educational Attainment. *Evolutionary Psychological Science*, *10*(3), 189–200. https://doi.org/10.1007/s40806-024-00395-2

Woodley, M. A., & Madison, G. (2013). Establishing an association between the Flynn effect and ability differentiation. *Personality and Individual Differences*, *55*(4), 387–390. https://doi.org/10.1016/j.paid.2013.03.016

Woodley of Menie, M. A., Peñaherrera-Aguirre, M., Fernandes, H. B. F., & Figueredo, A.-J. (2018). What Causes the Anti-Flynn Effect? A Data Synthesis and Analysis of Predictors. *Evolutionary Behavioral Sciences*, *12*(4), 276–295. https://doi.org/10.1037/ebs0000106

World Health Organization, & International Bank for Reconstruction and Development / The World Bank. (2023). *Tracking universal health coverage: 2023 global monitoring report. Executive summary*. World Health Organization. https://iris.who.int/bitstream/handle/10665/374060/9789240080393-eng.pdf

Wu, H., & Estabrook, R. (2016). Identification of Confirmatory Factor Analysis Models of Different Levels of Invariance for Ordered Categorical Outcomes. *Psychometrika*, *81*(4), 1014–1045. https://doi.org/10.1007/s11336-016-9506-0

Zainuddin, N. F., Jamaludin, M. N., Zulkapri, I., Mohd, N. A. R. N., & Haris, A. S. A. (2018). Investigating Variances Brain Activity of Individual Intelligence with Cycling Exercise: A Preliminary Study. In *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)* (pp. 408-412). IEEE.

Zhang, X., Chen, X. & Zhang, X. (2018). The impact of exposure to air pollution on cognitive performance. *Proceedings Of The National Academy Of Sciences*, *115*(37), 9193–9197. https://doi.org/10.1073/pnas.1809474115

Table of contents

Figures

Tables

Appendix A

Table A1

Systematic group differences between cohorts

Cohort	χ^2	*df*	*p*
2001 - 2011			
Education	22.154	4	< .001
Sex	5.694	1	.017
2001 - 2018			
Age	36.500	6	< .001
Sex	462.332	1	.026
2011 - 2018			
Age	29.669	6	< .001
Education	37.525	4	< .001
2011 - 2024			
Age	19.563	6	.003
Education	26.795	4	< .001
2018 - 2024			
Age	51.822	6	< .001

Note. Age has been grouped into age groups as follows: below 25; 26-35; 36-45; 46-55; 56-65; 66-75; 76+

Table A2

From MGCFA excluded items.

Items	*M*	*SD*
001	1.00	0.00
002	1.00	0.00
003	0.99	0.10
004	0.99	0.09
005	0.98	0.12
006	1.00	0.04
007	0.97	0.17
008	0.92	0.27
009	0.99	0.09
010	0.97	0.18
011	0.91	0.29
013	1.00	0.00
014	1.00	0.06
015	0.99	0.09
016	0.98	0.14
017	0.97	0.16
018	0.93	0.25
021	0.91	0.29
022	0.94	0.24
025	0.98	0.12
026	0.95	0.21
027	0.97	0.17
029	0.93	0.26
031	0.94	0.23
037	0.99	0.08
038	0.97	0.18
039	0.95	0.22
040	0.94	0.24
041	0.97	0.18

042	0.94	0.24

Appendix B

R Syntax – Data & Sociodemographics

```
# 0. Packages
install.packages("data.table", type = "binary")
install.packages("here")
install.packages("foreign", type = "binary")
install.packages("dplyr")
install.packages("effectsize")
install.packages("effsize")
install.packages("https://cran.r-project.org/src/contrib/Archive/lavaan/lavaan_0.6-18.tar.gz",
             repos = NULL, type = "source")
install.packages("parameters")

library(data.table)
library(here)
library(foreign)
library (dplyr)
library (effsize)
library(effectsize)
library(lavaan)
library(parameters)


# 1. Dataset vorbereiten

# 1.1 Datasets laden
spm_01 <-
as.data.table(read.spss(here("/Users/moritzfuchs/Desktop/Masterarbeit/Daten 2001-2018/SPM_2001.sav")))
```

```
spm_11 <-
as.data.table(read.spss(here("/Users/moritzfuchs/Desktop/Masterarbeit/Daten 2001-
2018/SPM_2011.sav")))

spm_18 <-
as.data.table(read.spss(here("/Users/moritzfuchs/Desktop/Masterarbeit/Daten 2001-
2018/SPM_2018.sav")))

spm_25 <-
as.data.table(read.spss(here("/Users/moritzfuchs/Desktop/Masterarbeit/Daten 2001-
2018/SPM_2025.sav")))

# 1.2 Variablennamen ändern (Großbuchstaben)

names(spm_01) <- toupper(names(spm_01))

# 1.3 Gesamt N pro Stichprobe

cohort_participants <- data.table("2001" = nrow(spm_01),
                                  "2011" = nrow(spm_11),
                                  "2018" = nrow(spm_18),
                                  "2025" = nrow(spm_25))


# 2. Sociodemographics

# 2.1 Geschlecht

sum(spm_01 == "männlich", na.rm = TRUE)
sum(spm_01 == "weiblich", na.rm = TRUE)

sum(spm_11 == "männlich", na.rm = TRUE)
sum(spm_11 == "weiblich", na.rm = TRUE)
```

```
sum(spm_18 == "männlich", na.rm = TRUE)
sum(spm_18 == "weiblich", na.rm = TRUE)

sum(spm_25 == "männlich", na.rm = TRUE)
sum(spm_25 == "weiblich", na.rm = TRUE)

# 2.2 Alter

round(mean(spm_01$AGE), 2)
round(sd(spm_01$AGE), 2)

round(mean(spm_11$AGE), 2)
round(sd(spm_11$AGE), 2)

round(mean(spm_18$AGE), 2)
round(sd(spm_18$AGE), 2)

round(mean(spm_25$AGE), 2)
round(sd(spm_25$AGE), 2)

# 2.3 Bildungsstand

# 2.3.1 Häufigkeit pro Jahrgang

table(spm_01$EDLEV, useNA = "ifany")
table(spm_11$EDLEV, useNA = "ifany")
table(spm_18$EDLEV, useNA = "ifany")
table(spm_25$EDLEV, useNA = "ifany")

# 2.3.2 Mittelwert

spm_01$EDLEV <- as.numeric(spm_01$EDLEV)
round(mean(spm_01$EDLEV), 2)
```

```
spm_11$EDLEV <- as.numeric(spm_11$EDLEV)
round(mean(spm_11$EDLEV), 2)

spm_18$EDLEV <- as.numeric(spm_18$EDLEV)
round(mean(spm_18$EDLEV), 2)

spm_25$EDLEV <- as.numeric(spm_25$EDLEV)
round(mean(spm_25$EDLEV), 2)

# 2.4 Chi2-Tests auf systematische Gruppenunterschiede

# 2.4.1 Alter

# Altersgruppen erstellen

breaks <- c(14, 26, 36, 46, 56, 66, 76, 90)
labels <- c("u25", "26-35", "36-45", "46-55", "56-65", "66-75", "76+")

spm_01$Agegroup <- cut(
  spm_01$AGE,
  breaks = breaks,
  labels = labels,
  right  = FALSE)

spm_11$Agegroup <- cut(
  spm_11$AGE,
  breaks = breaks,
  labels = labels,
  right  = FALSE)

spm_18$Agegroup <- cut(
  spm_18$AGE,
  breaks = breaks,
  labels = labels,
```

```
  right  = FALSE)

spm_25$Agegroup <- cut(
  spm_25$AGE,
  breaks = breaks,
  labels = labels,
  right  = FALSE)

# 2001 vs. 2025
tab <- table(Group    = factor(c(rep("2001", nrow(spm_01)),
                                  rep("2025", nrow(spm_25)))),
             Agegroup = c(spm_01$Agegroup, spm_25$Agegroup))
print(tab)
chi2_res <- chisq.test(tab, correct = TRUE)
print(chi2_res)

# 2001 vs. 2018
tab <- table(Group    = factor(c(rep("2001", nrow(spm_01)),
                                  rep("2018", nrow(spm_18)))),
             Agegroup = c(spm_01$Agegroup, spm_18$Agegroup))
print(tab)
chi2_res <- chisq.test(tab, correct = TRUE)
print(chi2_res)

# 2001 vs. 2011
tab <- table(Group    = factor(c(rep("2001", nrow(spm_01)),
                                  rep("2011", nrow(spm_11)))),
             Agegroup = c(spm_01$Agegroup, spm_11$Agegroup))
print(tab)
chi2_res <- chisq.test(tab, correct = TRUE)
print(chi2_res)

# 2011 vs. 2018
tab <- table(Group    = factor(c(rep("2011", nrow(spm_11)),
```

```
                         rep("2018", nrow(spm_18)))),
          Agegroup = c(spm_11$Agegroup, spm_18$Agegroup))
print(tab)
chi2_res <- chisq.test(tab, correct = TRUE)
print(chi2_res)

# 2011 vs. 2025
tab <- table(Group    = factor(c(rep("2011", nrow(spm_11)),
                         rep("2025", nrow(spm_25)))),
          Agegroup = c(spm_11$Agegroup, spm_25$Agegroup))
print(tab)
chi2_res <- chisq.test(tab, correct = TRUE)
print(chi2_res)

# 2018 vs. 2025
tab <- table(Group    = factor(c(rep("2018", nrow(spm_18)),
                         rep("2025", nrow(spm_25)))),
          Agegroup = c(spm_18$Agegroup, spm_25$Agegroup))
print(tab)
chi2_res <- chisq.test(tab, correct = TRUE)
print(chi2_res)

# 2.4.2 Bildungsstand

# 2001 vs. 2025
tab <- table(Group = factor(c(rep("2001", nrow(spm_01)),
                       rep("2025", nrow(spm_25)))),
          EDLEV  = c(spm_01$EDLEV, spm_25$EDLEV))
print(tab)
chi2_res <- chisq.test(tab, correct = TRUE)
print(chi2_res)

# 2001 vs. 2011
tab_01_11 <- table(Group = factor(c(rep("2001", nrow(spm_01)),
```

```
                          rep("2011", nrow(spm_11)))),
              EDLEV = c(spm_01$EDLEV, spm_11$EDLEV))
print(tab_01_11)
chi2_01_11 <- chisq.test(tab_01_11, correct = TRUE)
print(chi2_01_11)

# 2001 vs. 2018
tab_01_18 <- table(Group = factor(c(rep("2001", nrow(spm_01)),
                          rep("2018", nrow(spm_18)))),
              EDLEV = c(spm_01$EDLEV, spm_18$EDLEV))
print(tab_01_18)
chi2_01_18 <- chisq.test(tab_01_18, correct = TRUE)
print(chi2_01_18)

# 2011 vs. 2018
tab_11_18 <- table(Group = factor(c(rep("2011", nrow(spm_11)),
                          rep("2018", nrow(spm_18)))),
              EDLEV = c(spm_11$EDLEV, spm_18$EDLEV))
print(tab_11_18)
chi2_11_18 <- chisq.test(tab_11_18, correct = TRUE)
print(chi2_11_18)

# 2011 vs. 2025
tab_11_25 <- table(Group = factor(c(rep("2011", nrow(spm_11)),
                          rep("2025", nrow(spm_25)))),
              EDLEV = c(spm_11$EDLEV, spm_25$EDLEV))
print(tab_11_25)
chi2_11_25 <- chisq.test(tab_11_25, correct = TRUE)
print(chi2_11_25)

# 2018 vs. 2025
tab_18_25 <- table(Group = factor(c(rep("2018", nrow(spm_18)),
                          rep("2025", nrow(spm_25)))),
              EDLEV = c(spm_18$EDLEV, spm_25$EDLEV))
```

```
print(tab_18_25)
chi2_18_25 <- chisq.test(tab_18_25, correct = TRUE)
print(chi2_18_25)

# 2.4.3 Geschlecht

# 2001 vs. 2025
tab_01_25_sex <- table(Group = factor(c(rep("2001", nrow(spm_01)),
                                        rep("2025", nrow(spm_25)))),
                       SEX   = c(spm_01$SEX, spm_25$SEX))
print(tab_01_25_sex)
chi2_01_25_sex <- chisq.test(tab_01_25_sex, correct = TRUE)
print(chi2_01_25_sex)

# 2001 vs. 2011
tab_01_11_sex <- table(Group = factor(c(rep("2001", nrow(spm_01)),
                                        rep("2011", nrow(spm_11)))),
                       SEX   = c(spm_01$SEX, spm_11$SEX))
print(tab_01_11_sex)
chi2_01_11_sex <- chisq.test(tab_01_11_sex, correct = TRUE)
print(chi2_01_11_sex)

# 2001 vs. 2018
tab_01_18_sex <- table(Group = factor(c(rep("2001", nrow(spm_01)),
                                        rep("2018", nrow(spm_18)))),
                       SEX   = c(spm_01$SEX, spm_18$SEX))
print(tab_01_18_sex)
chi2_01_18_sex <- chisq.test(tab_01_18_sex, correct = TRUE)
print(chi2_01_18_sex)

# 2011 vs. 2018
tab_11_18_sex <- table(Group = factor(c(rep("2011", nrow(spm_11)),
                                        rep("2018", nrow(spm_18)))),
                       SEX   = c(spm_11$SEX, spm_18$SEX))
```

```
print(tab_11_18_sex)
chi2_11_18_sex <- chisq.test(tab_11_18_sex, correct = TRUE)
print(chi2_11_18_sex)

# 2011 vs. 2025
tab_11_25_sex <- table(Group = factor(c(rep("2011", nrow(spm_11)),
                                        rep("2025", nrow(spm_25)))),
                       SEX   = c(spm_11$SEX, spm_25$SEX))
print(tab_11_25_sex)
chi2_11_25_sex <- chisq.test(tab_11_25_sex, correct = TRUE)
print(chi2_11_25_sex)

# 2018 vs. 2025
tab_18_25_sex <- table(Group = factor(c(rep("2018", nrow(spm_18)),
                                        rep("2025", nrow(spm_25)))),
                       SEX   = c(spm_18$SEX, spm_25$SEX))
print(tab_18_25_sex)
chi2_18_25_sex <- chisq.test(tab_18_25_sex, correct = TRUE)
print(chi2_18_25_sex)
```

Appendix C

R Syntax – Cohen d & ANOVA

```
# 3. Gesamtscore

# 3.1 Mittelwert & SD

mean(spm_01$GS)
sd(spm_01$GS)

mean(spm_11$GS)
sd(spm_11$GS)

mean(spm_18$GS)
sd(spm_18$GS)

mean(spm_25$GS)
sd(spm_25$GS)

# 4. Effektstärken

# 4.1 Cohen d's Rohscores

cohen01vs11 <- cohen.d(spm_11$GS, spm_01$GS, na.rm = TRUE)$estimate
cohen01vs18 <- cohen.d(spm_18$GS, spm_01$GS, na.rm = TRUE)$estimate
cohen01vs25 <- cohen.d(spm_25$GS, spm_01$GS, na.rm = TRUE)$estimate
cohen11vs18 <- cohen.d(spm_18$GS, spm_11$GS, na.rm = TRUE)$estimate
cohen11vs25 <- cohen.d(spm_25$GS, spm_11$GS, na.rm = TRUE)$estimate
cohen18vs25 <- cohen.d(spm_25$GS, spm_18$GS, na.rm = TRUE)$estimate
```

```
# 4.2 t-test Rohscores

t.test(spm_11$GS, spm_01$GS, alternative = "two.sided")
t.test(spm_18$GS, spm_01$GS, alternative = "two.sided")
t.test(spm_25$GS, spm_01$GS, alternative = "two.sided")
t.test(spm_18$GS, spm_11$GS, alternative = "two.sided")
t.test(spm_25$GS, spm_11$GS, alternative = "two.sided")
t.test(spm_25$GS, spm_18$GS, alternative = "two.sided")


# 5. ANOVA

# 5.1 ANOVA gesamt


sum_table <- data.frame(Sum = c(spm_01$GS, spm_11$GS, spm_18$GS,
spm_25$GS),
                        Year = factor(rep(c("2001", "2011", "2018", "2025"),
                                 times = c(length(spm_01$GS),
                                           length(spm_11$GS),
                                           length(spm_18$GS),
                                           length(spm_25$GS)))))

result <- aov(Sum ~ Year, data = sum_table)
eta_squared(result)
summary(result)

# 5.2 ANOVA 01 vs. 11

sum_table <- data.frame(Sum = c(spm_01$GS, spm_11$GS),
                        Year = (rep(c(2001, 2011),
                                 times = c(length(spm_01$GS),
                                           length(spm_11$GS)))))
```

```
result <- aov(Sum ~ Year, data = sum_table)
eta_squared(result)
summary(result)

# 5.3 ANOVA 01 vs. 18

sum_table <- data.frame(Sum = c(spm_01$GS, spm_18$GS),
                    Year = (rep(c(2001, 2018),
                            times = c(length(spm_01$GS),
                                    length(spm_18$GS)))))

result <- aov(Sum ~ Year, data = sum_table)
eta_squared(result)
summary(result)

# 5.4 ANOVA 01 vs. 25

sum_table <- data.frame(Sum = c(spm_01$GS, spm_25$GS),
                    Year = rep(c(2001, 2025),
                            times = c(length(spm_01$GS),
                                    length(spm_25$GS))))

result <- aov(Sum ~ Year, data = sum_table)
eta_squared(result)
summary(result)

# 5.5 ANOVA 11 vs. 18

sum_table <- data.frame(Sum = c(spm_11$GS, spm_18$GS),
                    Year = rep(c(2011, 2018),
                            times = c(length(spm_11$GS),
                                    length(spm_18$GS))))

result <- aov(Sum ~ Year, data = sum_table)
```

```
eta_squared(result)
summary(result)

# 5.6 ANOVA 11 vs 25

sum_table <- data.frame(Sum = c(spm_11$GS, spm_25$GS),
                        Year = rep(c(2011, 2025),
                                   times = c(length(spm_11$GS),
                                             length(spm_25$GS))))

result <- aov(Sum ~ Year, data = sum_table)
eta_squared(result)
summary(result)

# 5.7 ANOVA 18 vs. 25

sum_table <- data.frame(Sum = c(spm_18$GS, spm_25$GS),
                        Year = rep(c(2018, 2025),
                                   times = c(length(spm_18$GS),
                                             length(spm_25$GS))))

result <- aov(Sum ~ Year, data = sum_table)
eta_squared(result)
summary(result)
```

Appendix D

R Syntax – MGCFA

```
# 6. Identifikation & Ausschluss problematischer Items

# 6.1 Itemscores definieren

itemscore_BEW <- sprintf("BEW%03d", 1:60)

# 6.2 Gesamt Means aller Items

item_means <- sapply(itemscore_BEW, function(it) {
  mean(c(
      spm_01[[it]],
      spm_11[[it]],
      spm_18[[it]],
      spm_25[[it]]),
    na.rm = TRUE)})

# 6.3 Gesamt SDs aller Items

item_sds <- sapply(itemscore_BEW, function(it) {
  sd(c(
      spm_01[[it]],
      spm_11[[it]],
      spm_18[[it]],
      spm_25[[it]]),
    na.rm = TRUE)})


item_stats_df <- data.frame(
  Item     = itemscore_BEW,
  Mean_All = round(item_means,  3),
```

```
  SD_All   = round(item_sds,    3),
  row.names = NULL)

print(item_stats_df)

# 6.4 Bereinigte Items

# 6.4.1 Overall

clean_items_table_overall <- data.frame(
  BEW012 = c(spm_01$BEW012, spm_11$BEW012, spm_18$BEW012,
spm_25$BEW012),
  BEW019 = c(spm_01$BEW019, spm_11$BEW019, spm_18$BEW019,
spm_25$BEW019),
  BEW020 = c(spm_01$BEW020, spm_11$BEW020, spm_18$BEW020,
spm_25$BEW020),
  BEW023 = c(spm_01$BEW023, spm_11$BEW023, spm_18$BEW023,
spm_25$BEW023),
  BEW024 = c(spm_01$BEW024, spm_11$BEW024, spm_18$BEW024,
spm_25$BEW024),
  BEW028 = c(spm_01$BEW028, spm_11$BEW028, spm_18$BEW028,
spm_25$BEW028),
  BEW030 = c(spm_01$BEW030, spm_11$BEW030, spm_18$BEW030,
spm_25$BEW030),
  BEW032 = c(spm_01$BEW032, spm_11$BEW032, spm_18$BEW032,
spm_25$BEW032),
  BEW033 = c(spm_01$BEW033, spm_11$BEW033, spm_18$BEW033,
spm_25$BEW033),
  BEW034 = c(spm_01$BEW034, spm_11$BEW034, spm_18$BEW034,
spm_25$BEW034),
  BEW035 = c(spm_01$BEW035, spm_11$BEW035, spm_18$BEW035,
spm_25$BEW035),
  BEW036 = c(spm_01$BEW036, spm_11$BEW036, spm_18$BEW036,
spm_25$BEW036),
```

```
  BEW043 = c(spm_01$BEW043, spm_11$BEW043, spm_18$BEW043,
spm_25$BEW043),
  BEW044 = c(spm_01$BEW044, spm_11$BEW044, spm_18$BEW044,
spm_25$BEW044),
  BEW045 = c(spm_01$BEW045, spm_11$BEW045, spm_18$BEW045,
spm_25$BEW045),
  BEW046 = c(spm_01$BEW046, spm_11$BEW046, spm_18$BEW046,
spm_25$BEW046),
  BEW047 = c(spm_01$BEW047, spm_11$BEW047, spm_18$BEW047,
spm_25$BEW047),
  BEW048 = c(spm_01$BEW048, spm_11$BEW048, spm_18$BEW048,
spm_25$BEW048),
  BEW049 = c(spm_01$BEW049, spm_11$BEW049, spm_18$BEW049,
spm_25$BEW049),
  BEW050 = c(spm_01$BEW050, spm_11$BEW050, spm_18$BEW050,
spm_25$BEW050),
  BEW051 = c(spm_01$BEW051, spm_11$BEW051, spm_18$BEW051,
spm_25$BEW051),
  BEW052 = c(spm_01$BEW052, spm_11$BEW052, spm_18$BEW052,
spm_25$BEW052),
  BEW053 = c(spm_01$BEW053, spm_11$BEW053, spm_18$BEW053,
spm_25$BEW053),
  BEW054 = c(spm_01$BEW054, spm_11$BEW054, spm_18$BEW054,
spm_25$BEW054),
  BEW055 = c(spm_01$BEW055, spm_11$BEW055, spm_18$BEW055,
spm_25$BEW055),
  BEW056 = c(spm_01$BEW056, spm_11$BEW056, spm_18$BEW056,
spm_25$BEW056),
  BEW057 = c(spm_01$BEW057, spm_11$BEW057, spm_18$BEW057,
spm_25$BEW057),
  BEW058 = c(spm_01$BEW058, spm_11$BEW058, spm_18$BEW058,
spm_25$BEW058),
  BEW059 = c(spm_01$BEW059, spm_11$BEW059, spm_18$BEW059,
spm_25$BEW059),
```

```
  BEW060 = c(spm_01$BEW060, spm_11$BEW060, spm_18$BEW060,
spm_25$BEW060),
  Year    = rep(
    c(2001, 2011, 2018, 2025),
    times = c(
      length(spm_01$GS),
      length(spm_11$GS),
      length(spm_18$GS),
      length(spm_25$GS))))

# 6.4.2 2001

clean_items_table_01 <- data.frame(

  BEW012 = spm_01$BEW012,
  BEW019 = spm_01$BEW019,
  BEW020 = spm_01$BEW020,
  BEW023 = spm_01$BEW023,
  BEW024 = spm_01$BEW024,
  BEW028 = spm_01$BEW028,
  BEW030 = spm_01$BEW030,
  BEW032 = spm_01$BEW032,
  BEW033 = spm_01$BEW033,
  BEW034 = spm_01$BEW034,
  BEW035 = spm_01$BEW035,
  BEW036 = spm_01$BEW036,
  BEW043 = spm_01$BEW043,
  BEW044 = spm_01$BEW044,
  BEW045 = spm_01$BEW045,
  BEW046 = spm_01$BEW046,
  BEW047 = spm_01$BEW047,
  BEW048 = spm_01$BEW048,
  BEW049 = spm_01$BEW049,
  BEW050 = spm_01$BEW050,
```

```
  BEW051 = spm_01$BEW051,
  BEW052 = spm_01$BEW052,
  BEW053 = spm_01$BEW053,
  BEW054 = spm_01$BEW054,
  BEW055 = spm_01$BEW055,
  BEW056 = spm_01$BEW056,
  BEW057 = spm_01$BEW057,
  BEW058 = spm_01$BEW058,
  BEW059 = spm_01$BEW059,
  BEW060 = spm_01$BEW060,
  Year   = rep(2001, times = length(spm_01$GS)))

# 6.4.3 2011

clean_items_table_11 <- data.frame(

  BEW012 = spm_11$BEW012,
  BEW019 = spm_11$BEW019,
  BEW020 = spm_11$BEW020,
  BEW023 = spm_11$BEW023,
  BEW024 = spm_11$BEW024,
  BEW028 = spm_11$BEW028,
  BEW030 = spm_11$BEW030,
  BEW032 = spm_11$BEW032,
  BEW033 = spm_11$BEW033,
  BEW034 = spm_11$BEW034,
  BEW035 = spm_11$BEW035,
  BEW036 = spm_11$BEW036,
  BEW043 = spm_11$BEW043,
  BEW044 = spm_11$BEW044,
  BEW045 = spm_11$BEW045,
  BEW046 = spm_11$BEW046,
  BEW047 = spm_11$BEW047,
  BEW048 = spm_11$BEW048,
```

```
  BEW049 = spm_11$BEW049,
  BEW050 = spm_11$BEW050,
  BEW051 = spm_11$BEW051,
  BEW052 = spm_11$BEW052,
  BEW053 = spm_11$BEW053,
  BEW054 = spm_11$BEW054,
  BEW055 = spm_11$BEW055,
  BEW056 = spm_11$BEW056,
  BEW057 = spm_11$BEW057,
  BEW058 = spm_11$BEW058,
  BEW059 = spm_11$BEW059,
  BEW060 = spm_11$BEW060,
  Year   = rep(2011, times = length(spm_11$GS)))

# 6.4.4 2018

clean_items_table_18 <- data.frame(

  BEW012 = spm_18$BEW012,
  BEW019 = spm_18$BEW019,
  BEW020 = spm_18$BEW020,
  BEW023 = spm_18$BEW023,
  BEW024 = spm_18$BEW024,
  BEW028 = spm_18$BEW028,
  BEW030 = spm_18$BEW030,
  BEW032 = spm_18$BEW032,
  BEW033 = spm_18$BEW033,
  BEW034 = spm_18$BEW034,
  BEW035 = spm_18$BEW035,
  BEW036 = spm_18$BEW036,
  BEW043 = spm_18$BEW043,
  BEW044 = spm_18$BEW044,
  BEW045 = spm_18$BEW045,
  BEW046 = spm_18$BEW046,
```

```
  BEW047 = spm_18$BEW047,
  BEW048 = spm_18$BEW048,
  BEW049 = spm_18$BEW049,
  BEW050 = spm_18$BEW050,
  BEW051 = spm_18$BEW051,
  BEW052 = spm_18$BEW052,
  BEW053 = spm_18$BEW053,
  BEW054 = spm_18$BEW054,
  BEW055 = spm_18$BEW055,
  BEW056 = spm_18$BEW056,
  BEW057 = spm_18$BEW057,
  BEW058 = spm_18$BEW058,
  BEW059 = spm_18$BEW059,
  BEW060 = spm_18$BEW060,
  Year   = rep(2018, times = length(spm_18$GS)))

# 6.4.5 2025

clean_items_table_25 <- data.frame(

  BEW012 = spm_25$BEW012,
  BEW019 = spm_25$BEW019,
  BEW020 = spm_25$BEW020,
  BEW023 = spm_25$BEW023,
  BEW024 = spm_25$BEW024,
  BEW028 = spm_25$BEW028,
  BEW030 = spm_25$BEW030,
  BEW032 = spm_25$BEW032,
  BEW033 = spm_25$BEW033,
  BEW034 = spm_25$BEW034,
  BEW035 = spm_25$BEW035,
  BEW036 = spm_25$BEW036,
  BEW043 = spm_25$BEW043,
  BEW044 = spm_25$BEW044,
```

```
  BEW045 = spm_25$BEW045,
  BEW046 = spm_25$BEW046,
  BEW047 = spm_25$BEW047,
  BEW048 = spm_25$BEW048,
  BEW049 = spm_25$BEW049,
  BEW050 = spm_25$BEW050,
  BEW051 = spm_25$BEW051,
  BEW052 = spm_25$BEW052,
  BEW053 = spm_25$BEW053,
  BEW054 = spm_25$BEW054,
  BEW055 = spm_25$BEW055,
  BEW056 = spm_25$BEW056,
  BEW057 = spm_25$BEW057,
  BEW058 = spm_25$BEW058,
  BEW059 = spm_25$BEW059,
  BEW060 = spm_25$BEW060,
  Year   = rep(2025, times = length(spm_25$GS)))

# 7. MGCFA

# 7.1 Modellspezifizierung

model <- 'reasoning =~ BEW012 + BEW019 + BEW020 + BEW023 + BEW024 +
BEW028 + BEW030 + BEW032 + BEW033 + BEW034 +
BEW035 + BEW036 + BEW043 + BEW044 + BEW045 + BEW046 + BEW047 +
BEW048 + BEW049 + BEW050 + BEW051 + BEW052 + BEW053 +
BEW054 + BEW055 + BEW056 + BEW057 + BEW058 + BEW059 + BEW060'

# 7.2 Measurment Invariance

# 7.2.1 Overall Fit

overall.fit <- cfa(model,
                   data = clean_items_table_overall,
```

```
                  estimator = "wlsmv",
                  std.lv = TRUE,
                  ordered = TRUE)
summary(overall.fit, standardized = TRUE, fit.measures=TRUE)

# 7.2.1 Overall Fit 2001

overall.fit01 <- cfa(model,
                    data = clean_items_table_01,
                    estimator = "wlsmv",
                    std.lv = TRUE,
                    ordered = TRUE)
summary(overall.fit01, standardized = TRUE, fit.measures=TRUE)

# 7.2.2 Overall Fit 2011

overall.fit11 <- cfa(model,
                    data = clean_items_table_11,
                    estimator = "wlsmv",
                    std.lv = TRUE,
                    ordered = TRUE)
summary(overall.fit11, standardized = TRUE, fit.measures=TRUE)

# 7.2.3 Overall Fit 2018

overall.fit18 <- cfa(model,
                    data = clean_items_table_18,
                    estimator = "wlsmv",
                    std.lv = TRUE,
                    ordered = TRUE)
summary(overall.fit18, standardized = TRUE, fit.measures=TRUE)

# 7.2.4 Overall Fit 2025
```

```
overall.fit25 <- cfa(model,
                  data = clean_items_table_25,
                  estimator = "wlsmv",
                  std.lv = TRUE,
                  ordered = TRUE)
summary(overall.fit25, standardized = TRUE, fit.measures=TRUE)

# 7.2.5 Configural

str(clean_items_table_overall)
clean_items_table_overall$Year <- as.factor(clean_items_table_overall$Year)

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

# 7.2.6 Strict

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

Appendix E

R Syntax – Latent Means & Latent Scores

```
# 8.1 Latent Means & Latent Sums

par <- parameterEstimates(residual_model.fit)

# 8.1.1 Ladungen pro Kohorte

par_01 <- subset(par, subset = (group == "1" & op == "=~"), select = c(rhs, est))
par_11 <- subset(par, subset = (group == "2" & op == "=~"), select = c(rhs, est))
par_18 <- subset(par, subset = (group == "3" & op == "=~"), select = c(rhs, est))
par_25 <- subset(par, subset = (group == "4" & op == "=~"), select = c(rhs, est))

spm01_clean <- as.data.table(clean_items_table_01)
spm11_clean <- as.data.table(clean_items_table_11)
spm18_clean <- as.data.table(clean_items_table_18)
spm25_clean <- as.data.table(clean_items_table_25)

# 2001
for (col_name in colnames(spm01_clean)) {
  if (col_name %in% par_01$rhs) {
    est_value <- par_01$est[par_01$rhs == col_name]
    spm01_clean[, (col_name) := get(col_name) * est_value]}}

# 2011
for (col_name in colnames(spm11_clean)) {
  if (col_name %in% par_11$rhs) {
    est_value <- par_11$est[par_11$rhs == col_name]
    spm11_clean[, (col_name) := get(col_name) * est_value]}}

# 2018
for (col_name in colnames(spm18_clean)) {
```

```
  if (col_name %in% par_18$rhs) {
    est_value <- par_18$est[par_18$rhs == col_name]
    spm18_clean[, (col_name) := get(col_name) * est_value]}}

# 2024
for (col_name in colnames(spm25_clean)) {
  if (col_name %in% par_25$rhs) {
    est_value <- par_25$est[par_25$rhs == col_name]
    spm25_clean[, (col_name) := get(col_name) * est_value]}}

# 8.2 Latent Sum-Scores

spm01_clean <- clean_items_table_01
spm11_clean <- clean_items_table_11
spm18_clean <- clean_items_table_18
spm25_clean <- clean_items_table_25

spm01_clean$SUM <- rowSums(spm01_clean[, c(
  "BEW012","BEW019","BEW020","BEW023","BEW024",
  "BEW028","BEW030","BEW032","BEW033","BEW034",
  "BEW035","BEW036","BEW043","BEW044","BEW045",
  "BEW046","BEW047","BEW048","BEW049","BEW050",
  "BEW051","BEW052","BEW053","BEW054","BEW055",
  "BEW056","BEW057","BEW058","BEW059","BEW060"
)], na.rm = TRUE)

spm11_clean$SUM <- rowSums(spm11_clean[, c(
  "BEW012","BEW019","BEW020","BEW023","BEW024",
  "BEW028","BEW030","BEW032","BEW033","BEW034",
  "BEW035","BEW036","BEW043","BEW044","BEW045",
  "BEW046","BEW047","BEW048","BEW049","BEW050",
  "BEW051","BEW052","BEW053","BEW054","BEW055",
  "BEW056","BEW057","BEW058","BEW059","BEW060"
)], na.rm = TRUE)
```

```
spm18_clean$SUM <- rowSums(spm18_clean[, c(
  "BEW012","BEW019","BEW020","BEW023","BEW024",
  "BEW028","BEW030","BEW032","BEW033","BEW034",
  "BEW035","BEW036","BEW043","BEW044","BEW045",
  "BEW046","BEW047","BEW048","BEW049","BEW050",
  "BEW051","BEW052","BEW053","BEW054","BEW055",
  "BEW056","BEW057","BEW058","BEW059","BEW060"
)], na.rm = TRUE)

spm25_clean$SUM <- rowSums(spm25_clean[, c(
  "BEW012","BEW019","BEW020","BEW023","BEW024",
  "BEW028","BEW030","BEW032","BEW033","BEW034",
  "BEW035","BEW036","BEW043","BEW044","BEW045",
  "BEW046","BEW047","BEW048","BEW049","BEW050",
  "BEW051","BEW052","BEW053","BEW054","BEW055",
  "BEW056","BEW057","BEW058","BEW059","BEW060"
)], na.rm = TRUE)

# 8.3 Latent Mean-Scores

spm01_clean$LM  <- rowMeans(spm01_clean[, c(
  "BEW012","BEW019","BEW020","BEW023","BEW024",
  "BEW028","BEW030","BEW032","BEW033","BEW034",
  "BEW035","BEW036","BEW043","BEW044","BEW045",
  "BEW046","BEW047","BEW048","BEW049","BEW050",
  "BEW051","BEW052","BEW053","BEW054","BEW055",
  "BEW056","BEW057","BEW058","BEW059","BEW060"
)], na.rm = TRUE)

spm11_clean$LM  <- rowMeans(spm11_clean[, c(
  "BEW012","BEW019","BEW020","BEW023","BEW024",
  "BEW028","BEW030","BEW032","BEW033","BEW034",
  "BEW035","BEW036","BEW043","BEW044","BEW045",
```

```
  "BEW046","BEW047","BEW048","BEW049","BEW050",
  "BEW051","BEW052","BEW053","BEW054","BEW055",
  "BEW056","BEW057","BEW058","BEW059","BEW060"
)], na.rm = TRUE)

spm18_clean$LM  <- rowMeans(spm18_clean[, c(
  "BEW012","BEW019","BEW020","BEW023","BEW024",
  "BEW028","BEW030","BEW032","BEW033","BEW034",
  "BEW035","BEW036","BEW043","BEW044","BEW045",
  "BEW046","BEW047","BEW048","BEW049","BEW050",
  "BEW051","BEW052","BEW053","BEW054","BEW055",
  "BEW056","BEW057","BEW058","BEW059","BEW060"
)], na.rm = TRUE)

spm25_clean$LM  <- rowMeans(spm25_clean[, c(
  "BEW012","BEW019","BEW020","BEW023","BEW024",
  "BEW028","BEW030","BEW032","BEW033","BEW034",
  "BEW035","BEW036","BEW043","BEW044","BEW045",
  "BEW046","BEW047","BEW048","BEW049","BEW050",
  "BEW051","BEW052","BEW053","BEW054","BEW055",
  "BEW056","BEW057","BEW058","BEW059","BEW060"
)], na.rm = TRUE)

# 8.4 Cohen's d für Latent Sum-Scores

cohen01vs11_sum <- cohen.d(spm11_clean$SUM, spm01_clean$SUM,
na.rm=TRUE)$estimate
cohen01vs18_sum <- cohen.d(spm18_clean$SUM, spm01_clean$SUM,
na.rm=TRUE)$estimate
cohen01vs25_sum <- cohen.d(spm25_clean$SUM, spm01_clean$SUM,
na.rm=TRUE)$estimate
cohen11vs18_sum <- cohen.d(spm18_clean$SUM, spm11_clean$SUM,
na.rm=TRUE)$estimate
```

```
cohen11vs25_sum <- cohen.d(spm25_clean$SUM, spm11_clean$SUM,
na.rm=TRUE)$estimate
cohen18vs25_sum <- cohen.d(spm25_clean$SUM, spm18_clean$SUM,
na.rm=TRUE)$estimate


# 8.5 Cohen’s d für Latent Mean-Scores

cohen01vs11_lm <- cohen.d(spm11_clean$LM,  spm01_clean$LM,  na.rm =
TRUE)$estimate
cohen01vs18_lm <- cohen.d(spm18_clean$LM,  spm01_clean$LM,  na.rm =
TRUE)$estimate
cohen01vs25_lm <- cohen.d(spm25_clean$LM,  spm01_clean$LM,  na.rm =
TRUE)$estimate
cohen11vs18_lm <- cohen.d(spm18_clean$LM,  spm11_clean$LM,  na.rm =
TRUE)$estimate
cohen11vs25_lm <- cohen.d(spm25_clean$LM,  spm11_clean$LM,  na.rm =
TRUE)$estimate
cohen18vs25_lm <- cohen.d(spm25_clean$LM,  spm18_clean$LM,  na.rm =
TRUE)$estimate

# 8.6 Effect Sizes

# 8.6.1 Latent Means

cat("Cohen's d (LM)\n",
   "2001 vs 2011:", round(cohen01vs11_lm,3), "\n",
   "2001 vs 2018:", round(cohen01vs18_lm,3), "\n",
   "2001 vs 2025:", round(cohen01vs25_lm,3), "\n",
   "2011 vs 2018:", round(cohen11vs18_lm,3), "\n",
   "2011 vs 2025:", round(cohen11vs25_lm,3), "\n",
   "2018 vs 2025:", round(cohen18vs25_lm,3), "\n\n")
```

```
# 8.6.2 Latent Sums

cat("Cohen's d (SUM)\n",
    "2001 vs 2011:", round(cohen01vs11_sum,3), "\n",
    "2001 vs 2018:", round(cohen01vs18_sum,3), "\n",
    "2001 vs 2025:", round(cohen01vs25_sum,3), "\n",
    "2011 vs 2018:", round(cohen11vs18_sum,3), "\n",
    "2011 vs 2025:", round(cohen11vs25_sum,3), "\n",
    "2018 vs 2025:", round(cohen18vs25_sum,3), "\n")

# 8.7 Signifikanztests

# 8.7.1 Latent Sums

t.test(spm01_clean$SUM, spm11_clean$SUM, alternative = "two.sided")
t.test(spm01_clean$SUM, spm18_clean$SUM, alternative = "two.sided")
t.test(spm01_clean$SUM, spm25_clean$SUM, alternative = "two.sided")
t.test(spm11_clean$SUM, spm18_clean$SUM, alternative = "two.sided")
t.test(spm11_clean$SUM, spm25_clean$SUM, alternative = "two.sided")
t.test(spm18_clean$SUM, spm25_clean$SUM, alternative = "two.sided")

# 8.7.2 Latent Means

t.test(spm01_clean$LM, spm11_clean$LM, alternative = "two.sided")
t.test(spm01_clean$LM, spm18_clean$LM, alternative = "two.sided")
t.test(spm01_clean$LM, spm25_clean$LM, alternative = "two.sided")
t.test(spm11_clean$LM, spm18_clean$LM, alternative = "two.sided")
t.test(spm11_clean$LM, spm25_clean$LM, alternative = "two.sided")
t.test(spm18_clean$LM, spm25_clean$LM, alternative = "two.sided")
```

Appendix F

R Syntax – KIs

```
install.packages("ggplot2")
install.packages("grid")

library(ggplot2)
library(grid)


# 9.KIs für die Plots

n01 <- length(spm_01$GS)
n11 <- length(spm_11$GS)
n18 <- length(spm_18$GS)
n25 <- length(spm_25$GS)


ci_d <- function(d, n1, n2, conf.level = 0.95) {
  df     <- n1 + n2 - 2
  se     <- sqrt((n1 + n2)/(n1*n2) + d^2/(2*(n1 + n2)))
  alpha  <- 1 - conf.level
  tcrit  <- qt(1 - alpha/2, df = df)
  lower  <- d - tcrit * se
  upper  <- d + tcrit * se
  c(d = d, CI_low = lower, CI_high = upper)}

# 9.1 KIs Rohscores

ci01_11 <- ci_d(cohen01vs11, n01, n11)
ci11_18 <- ci_d(cohen11vs18, n11, n18)
ci18_25 <- ci_d(cohen18vs25, n18, n25)
ci01_25 <- ci_d(cohen01vs25, n01, n25)
```

```
ci_table <- rbind(
  `2001–2011` = ci01_11,
  `2011–2018` = ci11_18,
  `2018–2025` = ci18_25,
  `2001–2025` = ci01_25)

print(ci_table)


# 9.2 KIs Latent Means

ci01_11_lm <- ci_d(cohen01vs11_lm, n01, n11)
ci11_18_lm <- ci_d(cohen11vs18_lm, n11, n18)
ci18_25_lm <- ci_d(cohen18vs25_lm, n18, n25)
ci01_25_lm <- ci_d(cohen01vs25_lm, n01, n25)


ci_table_lm <- rbind(
  `2001–2011` = ci01_11_lm,
  `2011–2018` = ci11_18_lm,
  `2018–2025` = ci18_25_lm,
  `2001–2025` = ci01_25_lm)

print(ci_table_lm)
```

Appendix G

R Syntax – Score Plot

```
install.packages("ggplot2")
install.packages("grid")

library(ggplot2)
library(grid)

# 10. Score Plot

summary_df <- data.frame(
  Year        = c(2001, 2011, 2018, 2024),
  Xpos        = 1:4,
  raw_mean    = c(
    mean(spm_01$GS,   na.rm = TRUE),
    mean(spm_11$GS,   na.rm = TRUE),
    mean(spm_18$GS,   na.rm = TRUE),
    mean(spm_25$GS,   na.rm = TRUE)),
  latent_mean = c(
    mean(spm01_clean$LM, na.rm = TRUE),
    mean(spm11_clean$LM, na.rm = TRUE),
    mean(spm18_clean$LM, na.rm = TRUE),
    mean(spm25_clean$LM, na.rm = TRUE)))

# 10.1 Label Daten
label_df <- summary_df
label_df$label   <- sprintf("Latent\nmean = %.2f", label_df$latent_mean)
label_df$nudge_x <- c( 0.08,  0.08,  0.08, 0.08)
label_df$nudge_y <- c( 1, 0, 0, 0)
label_df$size    <- 2.5
```

```
# 10.2 Plot
ggplot(summary_df, aes(x = Xpos, y = raw_mean)) +
  geom_vline(xintercept = summary_df$Xpos, color = "black", size = 0.6) +
  geom_line(color = "#E60073", size = 1.2) +
  geom_point(color = "#E60073", size = 4) +
  theme_minimal(base_family = "Arial", base_size = 12) +
  theme(panel.border       = element_rect(color = "black", fill = NA, size = 0.6),
        panel.grid.major.x  = element_blank(),
        panel.grid.minor    = element_blank(),
        panel.grid.major.y  = element_line(linetype = "dotted", color = "grey80"),
        axis.title.x       = element_text(size = 12),
        axis.title.y       = element_text(size = 12),
        axis.text.x        = element_text(size = 12),
        axis.text.y        = element_text(size = 12)) +


# 10.2.1 X- und Y-Achse
scale_x_continuous(
  breaks = summary_df$Xpos,
  labels = summary_df$Year,
  limits = c(1 - 0.15, 4 + 0.8),
  expand = c(0, 0)) +
scale_y_continuous(
  limits = c(45.5, 53.5),
  breaks = 46:53,
  expand = c(0, 0)) +

# 10.2.2 Beschriftungen
labs(
  x     = "Year",
  y     = "Raw means",
  title = "") +
```

```
# 10.2.3 Label-Boxen
geom_label(
  data       = label_df,
  aes(x = Xpos, y = raw_mean, label = label),
  fill       = "white",
  colour     = "#0099CC",
  label.size = 0.6,
  label.r    = unit(0.2, "lines"),
  size       = label_df$size,
  family     = "Arial",
  nudge_x    = label_df$nudge_x,
  nudge_y    = label_df$nudge_y,
  hjust      = 0,
  vjust      = 1)
```

Appendix H

R Syntax – Cohen d Plot

```
library(ggplot2)
library(grid)


plot_df <- data.frame(
  Comparison = c("2001 to 2011", "2011 to 2018", "2018 to 2024", "2001 to 2024"),
  Xpos       = 1:4,
  d_raw      = c(-0.07, 0.33, 0.47, 0.63),
  d_latent   = c(-0.14, 0.36, 0.44, 0.61),
  raw_lo     = c(-0.23, 0.17, 0.31, 0.45),
  raw_hi     = c( 0.09, 0.48, 0.63, 0.80),
  lat_lo     = c(-0.23, 0.17, 0.31, 0.44),
  lat_hi     = c( 0.09, 0.48, 0.63, 0.78))

# 11. Cohen d Plot
ggplot(plot_df, aes(x = Xpos)) +
geom_vline(xintercept = plot_df$Xpos, color = "black", size = 0.6) +
geom_errorbar(aes(ymin = raw_lo, ymax = raw_hi),
              width = 0.1, color = "#E60073", size = 0.8,
              position = position_nudge(x = -0.1)) +
geom_point(aes(y = d_raw),
            color = "#E60073", size = 6, shape = 21, fill = "#E60073",
            position = position_nudge(x = -0.1)) +

# Latent Cohen's d
geom_errorbar(aes(ymin = lat_lo, ymax = lat_hi),
              width = 0.1, color = "#0099CC", size = 0.8,
              position = position_nudge(x = 0.1)) +
geom_point(aes(y = d_latent),
```

```
        color = "#0099CC", size = 6, shape = 17,
        position = position_nudge(x = 0.1)) +

# DIQ-Boxen für Raw
geom_label(
  x         = 1 - 0.23,
  y         = plot_df$d_raw[1],
  label      = "DIQ = -1.05",
  fill       = "white",
  colour     = "#E60073",
  label.size = 0.4, label.r = unit(0.2, "lines"),
  size       = 2.5, family  = "Arial",
  hjust      = 1, vjust    = 0.5) +
geom_label(
  x         = 2 - 0.23,
  y         = plot_df$d_raw[2],
  label      = "DIQ = 7.07",
  fill       = "white",
  colour     = "#E60073",
  label.size = 0.4, label.r = unit(0.2, "lines"),
  size       = 2.5, family  = "Arial",
  hjust      = 1, vjust    = 0.5) +
geom_label(
  x         = 3 - 0.23,
  y         = plot_df$d_raw[3],
  label      = "DIQ = 11.75",
  fill       = "white",
  colour     = "#E60073",
  label.size = 0.4, label.r = unit(0.2, "lines"),
  size       = 2.5, family  = "Arial",
  hjust      = 1, vjust    = 0.5) +
geom_label(
  x         = 4 - 0.23,
  y         = plot_df$d_raw[4],
```

```
  label      = "DIQ = 4.11",
  fill       = "white",
  colour     = "#E60073",
  label.size = 0.4, label.r = unit(0.2, "lines"),
  size       = 2.5, family  = "Arial",
  hjust      = 1, vjust    = 0.5) +

# DIQ-Boxen für Latent
geom_label(
  x          = 1 + 0.23,
  y          = plot_df$d_latent[1],
  label      = "DIQ = -2.10",
  fill       = "white",
  colour     = "#0099CC",
  label.size = 0.4, label.r = unit(0.2, "lines"),
  size       = 2.5, family  = "Arial",
  hjust      = 0, vjust    = 0.5) +
geom_label(
  x          = 2 + 0.23,
  y          = plot_df$d_latent[2],
  label      = "DIQ = 7.71",
  fill       = "white",
  colour     = "#0099CC",
  label.size = 0.4, label.r = unit(0.2, "lines"),
  size       = 2.5, family  = "Arial",
  hjust      = 0, vjust    = 0.5) +
geom_label(
  x          = 3 + 0.23,
  y          = plot_df$d_latent[3],
  label      = "DIQ = 11.00",
  fill       = "white",
  colour     = "#0099CC",
  label.size = 0.4, label.r = unit(0.2, "lines"),
  size       = 2.5, family  = "Arial",
```

```
  hjust     = 0, vjust    = 0.5) +
geom_label(
  x         = 4 + 0.23,
  y         = plot_df$d_latent[4],
  label     = "DIQ = 3.98",
  fill      = "white",
  colour    = "#0099CC",
  label.size = 0.4, label.r = unit(0.2, "lines"),
  size      = 2.5, family  = "Arial",
  hjust     = 0, vjust    = 0.5) +
# Achsen
theme_minimal(base_family = "Arial", base_size = 12) +
  theme(panel.border       = element_rect(color = "black", fill = NA, size = 0.6),
        panel.grid.major.x  = element_blank(),
        panel.grid.minor    = element_blank(),
        panel.grid.major.y  = element_line(linetype = "dotted", color = "grey80"),
        axis.title.x       = element_text(size = 12),
        axis.title.y       = element_text(size = 12),
        axis.text.x        = element_text(size = 12),
        axis.text.y        = element_text(size = 12)) +
  scale_x_continuous(
    breaks = plot_df$Xpos,
    labels = plot_df$Comparison,
    limits = c(1 - 0.9, 4 + 0.9),
    expand = c(0, 0)) +
  scale_y_continuous(
    limits = c(min(plot_df$raw_lo, plot_df$lat_lo) - 0.1,
               max(plot_df$raw_hi, plot_df$lat_hi) + 0.1),
    breaks = scales::pretty_breaks(n = 7),
    expand = c(0, 0)) +
  labs(
    x = "",
    y = expression("Cohen " * italic(d)))
```

Appendix I

Abstract in English and German

Abstract

Positive generational IQ test score changes (i.e., the Flynn effect) have been globally observed for large parts of the 20th century. More recent research from the past decades, however, revealed quite ambiguous patterns, regarding strength and even direction of IQ trajectories across different countries and intelligence domains. Specifically, in the domain of fluid intelligence where, historically, gains have been most robust and largest, inconsistent findings seemingly point to a deceleration, stagnation or even a reversal of the Flynn effect. Here, we present measurement invariant data from four population-representative Germanophone cohorts' (total N = 1223) performance in an inductive reasoning task (Raven's SPM; Raven, et al., 2018) across four measurement points (2001, 2011, 2018 and 2024). Analyses revealed non-linear patterns. After performance stagnated between 2001 and 2011, positive changes could be documented subsequently. While some incremental changes between cohorts amounted to 11 IQ points per decade, the total detected Flynn effect over the observed time period from 2001 to 2024 amounted to changes of approximately 4 IQ points per decade, corresponding to a medium effect (Cohen d = 0.61). Our findings emphasize the importance of establishing measurement invariance in order to interpret Flynn effects properly. The implications of our results are being discussed in light of recently observed trends of domain-specific IQ trends and mechanisms of ability differentiation.

Zusammenfassung

Positive IQ-Testwertveränderungen über die Generationen hinweg (d.h. der Flynn Effekt) wurden für große Teile des 20. Jahrhunderts auf globaler Ebene beobachtet. Aktuelle Forschung der letzten Jahrzehnte zeigte jedoch uneindeutige Muster hinsichtlich der Stärke und sogar der Richtung der IQ-Verläufe über unterschiedliche Länder und Intelligenzdomänen hinweg. Insbesondere in der Domäne der fluiden Intelligenz, in der historisch betrachtet die robustesten und größten IQ-Gewinne stattfanden, deuten inkonsistente Befunde auf eine Verlangsamung, Stagnation oder sogar eine Umkehr des Flynn Effekts. Hiermit präsentieren wir messinvariante Daten einer populationsrepräsentativen deutschsprachigen Kohorte (totale N = 1223) in einem Test zum induktiven Denken (Raven's SPM; Raven et al., 2018) über vier Messzeitpunkte hinweg (2001, 2011, 2011 und 2024). Unsere Analyse ergab nicht-lineare Muster. Nachdem die Testleistung zwischen 2001 und 2011 stagnierte, konnten im weiteren Verlauf positive Veränderungen dokumentiert werden. Während manche inkrementelle Veränderungen zwischen den Kohorten 11 IQ-Punkte pro Dekade ausmachten, belief sich der gesamte Flynn Effekt über den untersuchten Zeitraum auf ungefähr 4 IQ-Punkte pro Dekade, was mit einem mittlerem Effekt einhergeht (Cohen d = 0.61). Unsere Befunde unterstreichen die Wichtigkeit Messinvarianz zu etablieren, um Flynn Effekte adäquat zu interpretieren. Die Implikationen unserer Ergebnisse werden in der Folge im Hinblick auf domänenspezifische IQ-Trends und Mechanismen der Fähigkeitsdifferenzierung diskutiert.