MASTERARBEIT | MASTER'S THESIS

Titel | Title

Cross-sectional assessment of the Flynn effect: evidence from the Grundintelligenztest Skala CFT 20

verfasst von | submitted by

Hannah Wainig BSc

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

Master of Science (MSc)

Wien | Vienna, 2024

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:	Ass.-Prof. Mag. Dr. Jakob Pietschnig Privatdoz.

Abstract

Generational intelligence gains (i.e., the Flynn effect) have been widely observed over the last century and show differences in countries, age and intelligence domains. However, in recent decades, observations of inconsistent patterns in form of stagnations or even reversals of intelligence development have been reported. The present paper investigates the German adoption of the Culture Fair Intelligence Test, the Grundintelligenztest Skala 20 and its revision CFT 20-R for possible Flynn effects on a healthy, German speaking sample (*N* = 157). Using a repeated measure ANOVA, results were differentiated according to intelligence domain when comparing IQ scores to the respective norms of 1977 and 2015. Figural reasoning (CFT) showed intelligence gains, but verbal (extension *Wortschatz*) and numerical abilities (extension *Zahlenfolge*) reported small intelligence losses. However, measurements of verbal and numerical IQ showed influences of ceiling effects. When considering this bias, results may then be interpreted as a stagnation or even small positive Flynn effect, suggesting a similar trend to that of figural intelligence. Therefore, present results of an above-average scoring sample indicate a stagnation or at least slowing of intelligence gains in Austria.

Keywords: Flynn effect, intelligence, Culture Fair Intelligence Test, Grundintelligenztest Skala 2

Zusammenfassung

Intelligenzzuwächse über Generationen (d.h., der Flynn-Effekt) sind im letzten Jahrhundert vielmals beobachtet worden und weisen Unterschiede zwischen Ländern, Altersgruppen und Intelligenzbereichen auf. In den letzten Jahrzehnten wurden jedoch inkonsistente Muster in Form von Stagnationen oder sogar Umkehrungen der Intelligenzentwicklung beobachtet. Die vorliegende Arbeit untersucht die deutsche Adaptation des Culture Fair Intelligence Test, die Grundintelligenztests Skala 20 und dessen Revision CFT 20-R auf mögliche Flynn-Effekte an einer gesunden, deutschsprachigen Stichprobe (N = 157). Unter Verwendung einer Repeated Measure ANOVA wurden IQ-Werte mit den jeweiligen Normen von 1977 und 2015 verglichen und differenzierte Ergebnisse hinsichtlich Intelligenz-Domäne gefunden. Figurales Denken (CFT) zeigte Intelligenzgewinne, aber verbale (Erweiterung Wortschatz) und numerische Fähigkeiten (Erweiterung Zahlenfolge) wiesen geringe Intelligenzverluste auf. Die Messungen des verbalen und numerischen IQ zeigten jedoch Einflüsse von Deckeneffekten. Wenn diese Verzerrung berücksichtigt wird, können die Ergebnisse als Stagnation oder sogar als leicht positiver Flynn-Effekt interpretiert werden, was auf einen ähnlichen Trend wie bei der figuralen Intelligenz hindeutet. Die vorliegenden Ergebnisse einer überdurchschnittlichen Stichprobe deuten also auf eine Stagnation oder zumindest Verlangsamung des Intelligenzzuwachses in Österreich hin.

Schlüsselwörter: Flynn Effekt, Intelligenz, Culture Fair Intelligence Test, Grundintelligenztest Skala 2

Introduction

It has been 40 years since James Flynn reported a rise in generational IQ test scores within the general population, introducing what is now called the Flynn effect (Flynn, 1984). Since then, numerous studies have tested this theory and found that while varying between countries, domains and years, the generational intelligence rise could be replicated various times over the past century (Pietschnig & Voracek, 2015). Flynn (2009) reported the strength of these IQ gains to be up to around 3 IQ points per decade, an analysis that was broadly confirmed by the extensive meta analysis of Pietschnig and Voracek (2015). However, the latter also reported that the changing trajectory did not occur in a linear fashion but has rather alternated between increasing and decreasing over the last century. Since 1995, reports of a stagnating or declining Flynn effect have continuously been found in Scandinavian countries (Flynn & Shayer, 2018) and have since then also been reported for various other European countries (e.g., Dutton et al., 2016).

In addition, it is important to note that the different intelligence domains have also developed to various degrees. It is commonly accepted that fluid intelligence (i.e., the ability to reason and solve tasks without prior knowledge; often tested through performance IQ tests) has increased more than crystallised IQ (i.e., knowledge based intelligence; often tested through verbal tests that cannot be solved by reasoning, such as vocabulary tests; Pietschnig & Voracek, 2015). Because of the recent stagnation or even negative development of the Flynn effect, it is now uncertain how various intelligence domains will develop. For example, a recent study by Lazaridis et al. (2022) found that while the Flynn effect was quite differentiated on domain level, it did not conform to the usually observed intelligence test score gains in crystallised and fluid IQ. Findings were ascribed to large differences within domains that have not been reported up to this point, as most studies before have used wider pooled domains focused on fluid and crystallised IQ instead of fine grained intelligence measures. Another explanation was provided in form of the impending stagnation or even reversal of intelligence gains.

To contribute to modern research on the Flynn effect, the present study will test these divergent findings further using the German adoption of the Culture Fair Test by Cattell (1960): the Grundintelligenztest Skala 20 (Weiß, 1978). The Culture Fair Test

(CFT) was developed as a measurement of general mental capacity, also referred to as factor g (Weiß, 1978). Factor g relies on the frequently observed concept that scores on different cognitive intelligence domains are positively correlated with each other within an individual creating a positive manifold of IQ test performance (Pietschnig & Voracek, 2015).

The CFT aims at testing this factor g without the influence of sociocultural, schooling or ethnic backgrounds (Weiß, 1978). To what extent a culture fair testing is truly possible has been questioned by various researchers. Some studies (e.g., Jensen, 1976) even argue that traditional standardised tests of intelligence such as Wechslers Intelligence Scale for Children, Raven's Progressive Matrices and Stanford-Binet show practically no evidence of cultural biases, hinting at an irrelevance of culture fair tests. Jensen (1976) further argues that the latter tests do not even measure factor g contrary to their declarations, but rather focus on specific intelligence domains. Moreover, some studies have also reported that the validity of culture fair tests is not higher than that of more traditional tests (Foreman, 1982). However, a recent study on the structural validity of the CFT by Troche et al. (2019) showed a higher validity than traditionally suggested when controlling for item-position effect. Item-position effect refers to the increasing influence that the processing of earlier items has on later items when homogenous items are provided. This effect was relevant to consider due to the structure of the CFT 20-R, which is divided into two homogenous parts.

Regarding the culture fairness of CFT 20, it is interesting that Weiß (1998) reports reduced intelligence scores for children with a migration background compared to their German peers. He ascribes this to a missing understanding of instructions, since, when Part 1 and 2 of the CFT 20 was applied, foreign children scored the same in Part 2 as their German peers of the same socioeconomic group. It therefore seems necessary that reservations of culture fair tests are considered when using them, especially in high relevance situations such as intelligence diagnostic or school enrollment (Foreman, 1982).

The German adoption of the Culture Fair Test is nonetheless one of the most widely used intelligence tests in Germanophone countries, allowing to test fluid intelligence as well as, through its extension of verbal and numerical testing ("Erweiterung Wortschatz und Zahlenfolge"), crystallised intelligence. The first German

adoption of the Grundintelligenztest Skala 2 (CFT 2) was developed by Weiß in 1972 and has since then been revised a handful of times. Due to the magnitude of revisions necessary for the CFT 2, Weiß published the CFT 20 in 1978 (using norms of 1977). Smaller changes were implemented for the most recent edition, the CFT 20-R, which was lastly normed in 2015. The two latter versions of Grundintelligenztest Skala 2 are relevant for the present paper.

An interesting point that stood out in our research, is a statement by the developer Dr. Weiß (2019) in the foreword of his most recent revision, the CFT 20-R. He states that:

> The so-called Flynn effect has been a much-heard buzzword in the past, which has led to irritation in some psychological advisory centers, among psycho-diagnosticians in the clinical field and in research projects. The analyses already reported in the first edition of the CFT 20-R, that the annual IQ increases claimed by Flynn are a statistical artefact, could be empirically proven once again and with even more extensive material through current control studies. For the fluid form of intelligence (named *Grundintelligenz* by me since 1971), there was no increase in the intelligence measured with the CFT 20-R in 2015 - i.e. over a period of 10-12 years! (p.10; Translated by author[1])

This is surprising since, as mentioned above, the Flynn effect has been widely accepted by the scientific community. Most commonly an IQ gain of 3 points per decade

[1] Der sog. Flynn Effekt war in der Vergangenheit ein vielgehörtes Schlagwort, das zu Irritationen in manchen psychologischen Beratungsstellen, bei Psychodiagnostikern im klinischen Bereich und bei Forschungsprojekten geführt hat. Die in der ersten Auflage des CFT 20-R bereits referierten Analysen, dass es sich bei von Flynn behaupteten jährlichen IQ-Steigerungen um ein statistisches Artefakt handelt, konnte erneut und mit noch umfangreicherem Material durch aktuelle Kontrolluntersuchungen empirisch belegt werden. Für die Fluid-Intelligenzform (von mir seit 1971 als Grundintelligenz benannt) gab es bei der 2015 für die mit dem CFT 20-R gemessene Intelligenz - also in einem Zeitraum von 10-12 Jahren - keine Steigerung! (Weiß, 2019, p.10)

and a stronger increase in fluid than in crystalline intelligence has been reported and replicated numerous times (Pietschnig & Voracek, 2015). However, Weiß (2019) found that, for example, CFT 20-R data of 2015 only varied slightly from the observed data of 1984 and thus reports no difference in fluid ability for a period of over 30 years.

The present paper therefore aims at examining these ambiguous findings closer, trying to answer the following two research questions:

(1) Does a Flynn effect exist in the CFT 20(-R)? In a cross sectional design, this would occur when participants scored higher in the CFT 20 (normed 1977) than in the CFT 20-R (normed 2015).

(2) If a Flynn effect exists, does it vary for different intelligence domains?

The latter is based on the research mentioned above in which it was commonly found that effects for fluid intelligence were stronger than for crystallised intelligence. Based on the considerations above, the following hypothesis were assumed:

Hypothesis 1: There will be significant score difference for the CFT 20/-R when calculating the IQ on the 1977 norms compared to the 2015 norms.

Hypothesis 2: There will be a significant score difference of the extension *Wortschatz* (WS) and *Zahlenfolge* (ZF) when calculating the IQ on the 1977 norms compared to the 2015 norms.

Hypothesis 3: The calculated Flynn effect will be higher for measures of fluid intelligence (CFT 20 vs. CFT 20-R and extension *Zahlenfolge* (ZF)) than for crystallised intelligence (extension *Wortschatz* (WS)).

The present paper will make contributions in several ways. The aim is to examine whether a Flynn effect is present in the CFT 20(-R), challenging to a degree the statement by Dr. Weiß (2019) that a Flynn effect is only a statistical artefact. Moreover, due to the recent change of trajectory in the Flynn effect, it is of interest to the present paper to study the current development of intelligence in Austria.

Theoretical background

The Flynn Effect

In an extensive meta analysis of the Flynn Effect, Pietschnig and Voracek (2015) combined results of 271 independent samples totaling nearly four million participants across 31 different countries from 1909 to 2013. As described above, the strength of global intelligence tests gains has been usually reported to around 3 IQ points per

decade (Flynn, 2009), a statistic that was broadly confirmed by the extensive research of Pietschnig and Voracek (2015). However, the latter reported that those gains of approximately 3 IQ points did not happen in a linear fashion. They found intelligence gains within the early 20th century to be relatively weak (0.80 IQ points per decade), followed by a severe increase in the 1920s (7.20 IQ points per decade), then a loss from 1935 to 1947 (2.10 IQ points per decade), but recovering until 1976 (3 IQ points per decade) and decreasing again to moderate gains of 2.30 IQ points per decade from then onwards. Finally, even more recent data suggests that since around 1995 a stagnation or even reversal of the Flynn effect has been occurring in Scandinavian countries (Flynn & Shayer, 2018), but has since then been also reported for various other European countries (e.g., Dutton et al., 2016). First signs of an impending reversal of the effect have recently even been found for North America (Dworak et al., 2023; Schroeder, 2019), which is surprising since most literature report a continuous positive Flynn effect for the United States (Flynn & Shayer, 2018). Thus, IQ gains seem to vary across time and countries, but were additionally also found to differ between age and intelligence domains. For example, various studies (for an overview, see Pietschnig & Voracek, 2015), found stronger gains were observed for adults than for children and larger increases for fluid intelligence than for crystallised IQ.

There are various possible influences that could have caused these intelligence changes. One of which could be historical events, such as can be seen above for World War I and II, when gains between the wars increased, but drastically decreased during World War II. Numerous other theories explaining the Flynn effect have also emerged in the last decades and were summarised by Pietschnig and Voracek (2015) into three groups: environmental, biological and hybrid (i.e., interaction of environmental and biological) factors.

Environmental factors

Education has been provided as one of the most named causes of the Flynn effect which would explain the global rise of crystallised intelligence. This is also in line with findings of larger intelligence gains in adults than in children, which has often been linked to better education. Though, if education would be the sole or main explanation, then crystallised IQ (which refers to knowledge based intelligence and is often tested through verbal tests that cannot be solved by reasoning; e.g., vocabulary tests) would be

expected to have risen more. However, it is widely accepted that fluid IQ (which refers to the ability to reason and solve tasks without prior knowledge; it is often tested through performance IQ tests) increased more than crystallised IQ (Flynn, 2009). Moreover, Pietschnig and Voracek (2015) reported no effects of age on crystallised and full-scale IQ (which refers to the average test score of various subtests examining various intelligence domains and is often calculated by averaging fluid and crystallised IQ test scores). It is therefore unlikely that education accounts for the Flynn effect by itself, but rather provides only a part of the explanation behind it.

Another environmental theory refers to the effects of technology. However, Pietschnig and Voracek (2015) found no conclusive evidence that individuals who frequently spend more time being exposed to visual media have increased fluid task performance. Nonetheless, modern technology is a possible factor that must not be entirely disregarded.

Third, a smaller family size has often been associated with a rise in intelligence. Findings that could not be replicated by the extensive meta analysis of Pietschnig and Voracek (2015), who suggest that family size most likely only plays a minimal role in the Flynn effect.

Lastly, test taking behaviour is discussed. Two mechanisms could be at play here. In the first half of the twentieth century, a severe increase in intelligence testing occurred in the Western world due to military and educational reasons, leading to people being more familiar to standardised tests and multiple-choice formats, which are commonly used in IQ tests. This test sophistication is related to feeling more comfortable with not only the format of IQ tests, but also to the person administering it; to allotting time more wisely or trying harder in test settings. Therefore, it is likely to have contributed to at least a small portion of intelligence gains. Since 1947 however, the cause of test sophistication is said to be relatively modest (Flynn, 2009). The second mechanism was proposed by Brand (1996), who proposed that changes in test taking behaviour, especially an increase in guessing, has resulted in higher test scores. This willingness to guess has been attributed to more lenient attitudes in Western societies. While evidence on this was provided by Must and Must (2012), intelligence gains have also been found in tests that do not use multiple choice formats (Williams, 2013). Therefore, while increased guessing behaviour may contribute to IQ gains, its role

appears limited, especially considering gains across various IQ domains and the historical trajectory of IQ increases. The data suggest that other factors beyond test-taking behaviour likely play a more significant role in driving IQ score changes over time (Pietschnig & Voracek, 2015).

Biological factor

One possible biological explanation is hybrid vigor, defined as the mating of genetically dissimilar individuals, aimed at increasing allelic heterozygosity and decreasing homozygosity. The theory posits that contemporary increases in individual mobility, leading to fewer consanguineous (i.e., marriages between individuals who share at least one common ancestor) and endogamous (i.e., marriage within one's own tribe or group) marriages, might explain IQ gains observed in recent decades (Mingroni, 2007). However, despite offering a plausible mechanism, the relevance of hybrid vigor is constrained when considering the magnitude and pace of IQ gains. Evaluations of the proposed model suggest that, even under optimal conditions, hybrid vigor may only account for a portion of observed IQ gains, falling short of explaining the substantial gains observed globally over the past century (Pietschnig & Voracek, 2015).

Hybrid factors

In their meta analysis, Pietschnig and Voracek (2015) propose numerous hybrid factors, including more basic background factors, such as blood lead levels, genomic imprinting, nutrition, and reduced pathogen stress. For example, improvements in nutrition have been associated with stronger cognitive development, coinciding with observed IQ test score gains over time and research suggests improved prenatal and postnatal nutrition may explain part of the Flynn effect, where IQ scores have risen over time (Lynn, 2009). More complex, integrative theories have also been introduced. These include reduced IQ variability, effects of social multipliers, and decreasing life history speed (for a complete overview see Pietschnig & Voracek, 2015).

Recent development of the Flynn effect

The causes provided above could in turn also explain the stagnation of the Flynn effect in some countries, since it is possible that the factors have lost their strength due to ceiling effects (for example, increased test guessing can on average only achieve higher scores up to a limited extent) or diminishing returns (for example, longer schooling; Pietschnig et al., 2018). At least in Western nations, advancements in

nutrition, healthcare, and sanitation may have plateaued in recent decades, resulting in a stagnation of developmental test score improvements (Pietschnig et al., 2021).

Further explanations for the apparent decrease of intelligence have been offered. Lynn (2011) reports fertility as a possible cause, as higher fertility should be related to lower test scores due to selective population reproduction. Dutton and Lynn (2013) mention migration, as people with lower intelligence migrate to nations with higher IQs resulting in declining scores in those countries. Lastly, mortality could also be a reason due to the advancement of modern medicine possibly resulting in an increased likelihood of less physically fit individuals reaching reproductive age. This trend could contribute to a decline in population cognitive ability due to the positive association between physical and psychological fitness (Nyborg, 2012). However, a meta analysis by Pietschnig et al. (2018) showed that neither of those three reasons appear to have significant influence on the changing IQ trajectory. Instead, they focus on the explanation of a negative relation of psychometric *g* (i.e., the general cognitive ability of individuals) to test score changes. In the past years, there has been a trend to specialise in certain abilities, for example in the work setting. Pietschnig et al. (2023) explain the repercussions of this on intelligence test scores using the analogy of a decathlon. If a decathlete were to focus on only one discipline, they would improve and gain more points in this field. This would in turn increase their overall score as long as the other subdisciplines would not decrease a substantial amount and/or a limit within the trained discipline would not be reached. Transferring this analogy to intelligence, we can assume that once a ceiling is reached, this should result in a decrease of *g* and in turn result in a decreasing IQ score (Pietschnig et al, 2023). Additionally, an increase in specific abilities may have masked g-based ability declines before the above-mentioned ceiling was reached (Pietschnig et al., 2018).

Differences in changes between intelligence domains were also found by Lazaridis et al. (2022) using the Cattell-Horn-Carroll (CHC) intelligence model. The CHC model stands as one of the most widely embraced comprehensive frameworks for understanding human intelligence and includes measures for various types of intelligence from fluid reasoning to learning efficiency and quantitative knowledge. Lazaridis et al. (2022) researched 10 stratum II domains from 1996 to 2018 in 36 mostly population representative Germanophone standardisation samples and found

differentiating Flynn effect results between intelligence domains. However, they could not replicate the findings of test score gains in crystallised and fluid intelligence. While positive changes were somewhat prevailing, the effects varied in strength and across domains. Clear positive and significant changes were observed for the full-scale IQ (g), comprehension-knowledge, learning- efficiency and general domain-specific knowledge. No test score changes were found for processing speed. Reading, writing and memory capacity showed a moderate negative effect.

Surprisingly, most investigated domains showed ambiguous developments between different test instruments (Lazaridis et al., 2022), which was particularly unexpected for fluid reasoning since it has been repeatedly reported to show the highest gains in comparison to other intelligence domains (Flynn, 2009). Lazaridis et al. (2022) provide several reasons for the ambiguous results of fluid intelligence including firstly, the detailed analysis of specific abilities compared to the usually wider pooled domains resulting in more differentiated changes. Further evidence on this is provided by Oberleiter et al. (2024), who assume, based on their own findings, that the conflicting research on Flynn effect might be at least partly due to the relatively coarse assessments of intelligence in previous literature and hence urge the usage of more fine-grained intelligence measures.

Another reason provided by Lazaridis et al. (2022) is the possibility that Germanophone populations simply differ from other populations, which seems plausible since numerous studies such as Pietschnig and Voracek (2015) and Pietschnig et al. (2010) have found significant changes between countries in regard to the Flynn effect. However, even if this were the case, this alone could not explain the changing trajectory within German speaking populations, since over most of the 1900s significant positive changes were found for this population. Thus Lazaridis et al. (2022) circle back to an important possible cause: the non-linearity of the Flynn effect, which has gained evidence in the last decades (Pietschnig & Voracek, 2015). Flynn effect findings are commonly presented as linear changes over decades, which is due to the inherent design characteristics of studies investigating the Flynn effect that typically employ decadal assessments (Lazaridis et al., 2022). This could mean that differences on the same test are due to genuine changes of strength or direction of the Flynn effect because of various influences such as world wars (Pietschnig & Voracek, 2015) or

educational reforms (Baker et al., 2015). However, Lazaridis et al. (2022) argue that non-linearity alone is also inapt to explain their reported vast changes and urge to consider a combination of their mentioned causes (for a full overview, see Lazaridis et al., 2022) as well as the possibility of an imminent stagnation or even reversal of the Flynn effect.

At this point, it is important to note a relevant challenge when assessing the Flynn effect, which revolves around discerning whether alterations in test scores truly reflect shifts in the population's abilities or simply signify manifestations of differential item functioning (DIF) across various assessment years (Oberleiter et al., 2024). DIF refers to the phenomenon in which differences in average performance between samples arise from variations in item difficulty or their capacity to discriminate between levels of ability, rather than genuine disparities in abilities as societal norms and cultural perceptions evolve. This variation leads individuals to approach tests with differing levels of knowledge, thereby influencing the perceived difficulty of specific items (Gonthier & Grégoire, 2022). For instance, a study by Dutton and Lynn (2015) found effects for a declining IQ in France, but their findings were questioned a couple of years later by Gonthier et al. (2021). The latter found evidence that the presented negative Flynn effects cannot be ascribed to a real intelligence decline but were rather the consequence of test designs that lacked measurement invariance. Measurement invariance refers to the association between items (or test scores) and latent factors (or traits) not depending on group membership or time points (Mellenbergh, 1989).

The examination of measurement invariance is only possible for IQ tests based on item response theory (IRT) as only then it is possible to directly examine whether all items are related to the same latent ability and thus compare group scores on the same scale. However, most studies of the Flynn effect are based on tests following the classical test theory (CTT) approach (Pietschnig et al., 2013). Therefore, meaningful interpretation of test score changes as real changes in population ability requires the establishment of cross-temporal measurement invariance, indicating the absence of DIF and the constancy of item properties over time. A study fulfilling these requirements was conducted by Oberleiter et al. (2024) providing evidence for a real negative Flynn effect, at least for fluid intelligence. However, the extent to which the ambiguous patterns of the

Flynn effect could be better explained by item drift or domain specificity still remains unclear.

To contribute to the examination of the Flynn effect, the present paper utilised the widely used Grundintelligenztest Skala 20 (Weiß, 1978).

Grundintelligenztest Skala CFT 20(-R)

One of the most widely used intelligence tests in Germanophone countries is the Grundintelligenztest Skala CFT (Weiß, 1972), a German adoption of the Culture Fair Intelligence Test by R.B. Cattell (1960), that allows testing of fluid and crystallised intelligence through three parts: the CFT for figural reasoning, an extension for dyslexia (*Wortschatz*, WS) and an extension for numerical sequences (*Zahlenfolge*, ZF). For ages 5.3 to 9.5 and 6.6 to 11.11 the Grundintelligenztest Skala 1 (CFT 1) (Weiß & Osterman 2013) and for ages 8.5 to 19.11 and adults from 20 to 64 years the Grundintelligenztest Skala 2 (CFT 2) was developed (Weiß, 1972). The latter is of interest in the present study. It was profoundly refined and published under the name CFT 20 in 1978 due to various suggestions including norming differences. Between then and the revision CFT 20-R (Weiß, 2006) numerous examinations took place to test actuality of norms and found that when controlling for age and increased percentage of foreigners, the norms of 1977 stayed relevant and did not have to be changed. However, the CFT 20 was again revised in the mid 2000s by Weiß (2006) including new norms and then retested in 2015, slightly altering the norms to the current standards relevant today (Weiß, 2019). When comparing scores between CFT 20 (normed 1977) and CFT 20-R (using 2003 norms of the first edition), Weiß (2019) found no confirmation of a Flynn effect in the proposed strength. No concrete numbers were presented in the context of comparing for ages above 20 years, but norms were reportedly adapted between first to second edition of the CFT 20-R (normed 2015; Weiß, 2019).

A study by Hagmann-von Arx et al. (2018) compared various intelligence tests, including the CFT 20-R, and categorised the tests by standardisation year, with the CFT 20-R representing one of the older procedures. The study observed that children scored lower on more recently standardised tests than in older measurements, providing serious evidence for a positive intelligence development over the years. These contradicting findings around the Flynn effect and the Grundintelligenztest Skala are examined further in the present paper.

Method

Before accessing any data, the study design, analysis plan and specific main study hypotheses were pre-registered on the Open Science Framework (OSF; https://osf.io/2utjc, registered and last edited on 22 February 2023).

Participants

In total, a number of 157 participants were recruited for the present study (63.7% women; age: *M* = 29.27, *SD* = 13.99, range: 18-79 years). Requirements were a minimum age of 18 years, fluency in German and no disabilities concerning cognitive function. Probands were recruited via social media (Facebook, Instagram, WhatsApp) through the circle of acquaintances of the author and participation was voluntary and anonymous. Subjects were naturally allowed to terminate the testing prematurely, however, since no participant opted to do so, no exclusions from the analysis were necessary.

Materials

To assess participants' intelligence scores the CFT 20 and CFT 20-R as well as the extensions *Wortschatz* (WS) and *Zahlenfolge* (ZF) were used. The test is based on the General Fluid Ability, which refers to the ability of recognising complex relationships and applying them in novel situations. To also test General Crystallised Ability, which Weiß (2007) defines according to Cattell as general knowledge, vocabulary, language comprehension as well as handling numbers, an extension of dyslexia (*Wortschatz*, WS) and numerical sequence (*Zahlenfolge*, ZF) was added. This addition was important since Jäger (1982) reported that in the intelligence model of Cattell, the *processing capacity* factor (“Verarbeitungskapazität”) on the operational dimension includes three dimensions: a figural dimension (which can be tested through the CFT) as well as a verbal and numerical dimension. Though while Weiß (2007) attributes ZF to crystallised intelligence (with a component of fluid intelligence), the present paper allocates it towards fluid intelligence (with a component of crystallised intelligence), since other studies (e.g., Lazaridis et al., 2022) and tests (e.g., I-S-T 2000R; Liepmann et al., 2007) using numerical sequence, also categorised it within fluid intelligence. Moreover, ZF only requires basic arithmetic operations (adding, subtracting, multiplying, and dividing) of small numbers and no sophisticated calculations are necessary (Weiß, 2007). While the knowledge of the basic arithmetic operations could be based towards crystalline

intelligence, it can be expected for at least the participants of the present sample (who are all above the age of 18) that these basic operations are so ingrained within the individuals that the crystallised intelligence only plays a small role. It seems more important for the completion of the task to quickly recognise the relationships between the numbers and applying this concept to the next missing number. In the understanding of the present study, it does therefore require more fluid than crystallised intelligence.

In either way, the extension makes it possible to test the full-scale IQ by testing both fluid and crystallised intelligence. As described in the manual, the full-scale IQ is calculated by averaging the three IQ scores to equal parts (Weiß, 2007).

Changes in item count and difficulty occurred between the CFT 20 and its revision. In total, a number of 29 items spread out over all subtests differed between CFT 20 and CFT 20-R (CFT 20 tests 92 items in total; CFT 20-R tests 101 items). 10 items were only offered in the CFT 20 and 19 items were solely tested in the CFT 20-R; all other items were overlapping. Weiß (2019) reports the changed items of the revision to be of a higher difficulty level.

Therefore, when analysing the actuality of norms, Weiß (2019) used percentages of solved items instead of raw scores to accurately compare results. Results of CFT 20 and CFT 20-R for ages 9 to 19 were examined. No significant differences were found, except for ages 17.1 to 19, which increased from 67.0 to 70.5 percent in Part 1. For Part 2 a decrease from 69.7 to 67.8 was found from CFT 20 to CFT 20-R in the same age group. Totaling this up, only a very slight increase from CFT 20 to CFT 20-R was found for ages 17.1 to 19 of 68.3 to 68.9 respectively. He furthermore reports that when considering the varying difficulty levels of items and excluding items that were more complex in the revision than in the original CFT 20, the courses of curves stayed the same, supporting his argument that no Flynn effect is present for the Grundintelligenztest Skala.

CFT 20/-R

The CFT 20/-R consists of two parallel parts of four subtests each. For the current study, the whole test (i.e., both parallel parts) was administered. Reliability of the CFT 20 and CFT 20-R is reported to be $\alpha = .95$ and $\alpha = .96$ respectively. Correlations between part one and two are $r = .80$ (1977) for CFT 20 and $r = .82$ (2003) for CFT 20-R (also nearly all internal correlations are reported higher for the revision; Weiß, 2019). The four

subtests are Series, Classifications, Matrices and Topologies, together testing figural reasoning. In the Series subtest, participants are presented with a sequence of three figures and are required to select the fourth figure from a set of five alternatives that continues the progressive change observed in the series. In the Classifications subtest, each item features five figures, with one figure differing from the rest based on a specific feature, such as orientation. Participants must identify the deviant figure among the set. In the Matrices subtest, items consist of either a 2x2 or a 3x3 matrix, with the bottom-right cell left empty. Participants are tasked with identifying the figure that best completes the pattern in the matrix, choosing from five alternatives. In the Topologies subtest, participants are presented with a reference configuration of geometric figures, along with one or more dots. From five alternatives, participants must select a configuration where the dot(s) exhibit the same topological relationship to the parts of the configuration as in the reference configuration. A study concerning the structural validity of the CFT found that these four subtests can be split into two types of latent variables being tested, one latent variable for Series and Matrices, one for Classifications and Topologies. The variables were found to be substantially correlated with each other, thus making it possible to obtain a second-order factor of inductive reasoning (Troche et al., 2016).

To test both CFT 20 and its revision on all participants, the two tests were combined. Originally, both tests consist of two forms (Form A and B) that only slightly differ in sequence of options provided within an item. Weiß (2019) reports these forms as completely equivalent to each other, testing the same or equal items in a different order (i.e., *pseudo parallel forms*), which is why only Form A of both CFT 20 and CFT 20-R were combined for the examination of this study. While most items were identical between CFT 20 and CFT 20-R, a maximum of three items per subtest differed between them. In Part 1 of the original CFT 20 no items were extracted for the revision, but items with a higher difficulty level were added. In Part 2, some items were exchanged between CFT 20 and CFT 20-R. Items with a higher complexity level were introduced to the revision while simultaneously removing some items with a lower difficulty level. In both tests participants first completed two to three trial items to grasp the concept of each subtest and then the difficulty level increased with each item (Weiß, 2006).

To counterbalance a possible influence of item order, two different test versions had to be compiled: One started with items of CFT 20 and added the additional items of the CFT 20-R at the end of each subscale respectively. The second test version started with items of CFT 20-R and ended with the non-overlapping items of CFT 20. This resulted in 11 to 15 items per subtest for both Part 1 and 2 (compared to 8 to 15 items in the original test manuals). It is important to note that especially in the case of Part 2 of the test version with the CFT 20-R as basis, the added items of the CFT 20 were at the end of the test. As mentioned above, the latter items are according to Weiß (2019) mostly less difficult. If this assumption is correct, then a partly atypical progression of difficulty could be the result. Since the CFT is a time sensitive test, this could have repercussions on individuals being able to try and solve all provided items.

Extension Wortschatz and Zahlenfolge

The extension remained the same over the course of CFT 20 and CFT 20-R for both Wortschatz (WS) and Zahlenfolge (ZF), thus no changes had to be made. Only one change was made to one item of the vocabulary test of the revision to adhere to the new orthography (“neue Rechtschreibung”). This later version was used for all participants as the new orthography has been exhaustively implemented in Austria since the mid 2000s and it can be assumed that even older participants have adapted to it by the present date (nearly 20 years later). The extension consists of Form A and B in both cases, which were both used. Form A and B were constructed as pseudo parallel forms, testing the same or equal items in a deviating order to prevent copying off other participants (Weiß, 2007),

Wortschatz aimed at testing beyond the German core vocabulary, but within conversational language; thus assessing verbal intelligence as well as general education. It consists of 30 multiple choice items, each holding one key word to which the word with the same or most similar meaning had to be chosen out of five (for example, key word “Bluse” (blouse) and options of “Hemd” (shirt), “Wind” (wind), “Anzug” (suit), “Apparat” (device), “Stärke” (strength). The first option would be correct).

Zahlenfolge aimed at recognising rules and regularities of simple to more complex numerical tasks. It consists of 21 multiple choice items, each containing a numerical sequence with six numbers to which the seventh number had to be chosen

out of five possible options (for example, number sequence of “10, 15, 20, 25, 30, 35, ?” and options “38”, “45”, “42”, “40”, “41”; of which “40” would be correct).

Procedure

The combined versions of CFT 20 and CFT 20-R as well as the unmodified extension were administered in pen and paper format on 157 participants. Weiß (2019) recommends using both Part 1 and 2 of the CFT (i.e., long form) for higher reliability and validity, which we adhered to. Testing took place in group settings between April and November of 2023 and minimum time was given (maximum time is only recommended for people with cognitive disabilities or in special diagnostic settings). Participants had four minutes to solve Subtest 1 and 2 of Part 1 of the CFT and three minutes for all other subtests of the CFT. 12 minutes were each provided for Wortschatz and Zahlenfolge while instruction took approximately 10 minutes, totaling to approximately 55 minutes. Participants were randomly assigned to either of the two test versions assembled by the author of the present paper as well as to Form A or B of the extension. If participants received Form A for Wortschatz, they also received Form A for Zahlenfolge; however, this was not aligned to the test version of CFT they received (i.e., participants who received test version 1 could receive Form A or B of the extension and the same is true for test version 2), ensuring a truly random distribution.

Analysis

Data analysis was conducted using IBM SPSS v. 29 and consisted of four main steps. For all analytical steps a significance level of $p < .05$ was predefined.

Data Transformation

First, raw scores of every participant in each subtest were transformed into IQ scores using two different sets of norms. On one hand, data was transformed using the norms of 1977, which are applicable for the CFT 20, on the other hand the norms of 2015 for CFT 20-R were employed. The mean age of our participants (N = 157) was 29.27 (SD = 13.992), hence age norms of 20-29 years were used for the CFT and the mean norms of ages 20-24 and 25-29 were used for the revision to create equal age brackets. For the latter, the two IQ scores for each age and raw score were averaged; for example, a raw score of 41 is equivalent to an IQ score of 104 for ages 20-24 and 106 for ages 25-29. Thus, if a participant scored 41 points, this would equate to 105 IQ points in this sample. For the extension, norms of age 13,7 to 15,6 had to be used, as these were the highest

age norms reported for the WS and ZF 20. For its revision the mean norms for the ages 13,1-14,0, 14,1-15,0 and 15,1-16,0 had to be used for similar age frames and were calculated as described above for CFT. The subtest IQ scores were used to determine an IQ score for CFT at T1 (meaning compared with norms of 1977) and T2 (meaning compared with norms of 2015), Wortschatz old and new as well as Zahlenfolge old and new. The three different aspects of intelligence were then averaged to calculate a full-scale IQ T1 and full-scale IQ T2 for each participant according to the manual.

First ANOVA

Firstly, a repeated measure ANOVA was performed to test whether the full-scale IQ scores differed in terms of standardisation year (1977 vs. 2015). The latter was used as within-subject factor.

Second ANOVA

A two way repeated measure ANOVA was performed to examine the development of each intelligence domain over time. IQ results were split according to intelligence test (CFT, WS and ZF) and then each domain was compared to its results calculated with norms of T1 versus T2 (i.e., standardisation year: 1977 vs. 2015). Standardisation year and type of test were both used as within-subject factor.

Item Difficulty

Lastly, item difficulty between items that were changed from CFT 20 to version CFT 20-R had to be explored. To do so, item raw scores are brought on to the same scale by multiplying the scores of the changed items for each subscale with the numerical differences of changed items. This allowed scores to be compared via a paired t-test to measure whether participants solved an equal number of exchanged items. If no significant difference is found, the not-overlapping items are considered equally difficult.

In case of the Grundintelligenztest Skala by Weiß (2006), a total of 10 items were added to Part 1 of the CFT 20-R (1 item was added to Subtest 2 and 3 items were added to the remaining subtests), while keeping all items of CFT 20. For Part 2 of the CFT 20-R, 10 items of the CFT 20 were removed and exchanged for 9 other items (2 items were exchanged in Subtest 1; 3 items were removed and 1 item added in Subtest 2; 3 items were exchanged in Subtest 3; 2 items were removed and 3 items added in Subtest 4). In the combined test versions, all items (overlapping as well as not-

overlapping) were administered to the participants. Therefore, relative raw scores of the not-overlapping items of Part 2 for each subtest respectively could be calculated. Items were brought on to the same scale as described above by multiplying scores with the numerical difference of changed items. For example, in Subtest 2 three items were only present in the CFT and one item was only present in the CFT 20-R, so the score of the three items is multiplied by one divided through three (1/3). Therefore, a score of 3 points for these three not-overlapping items is now equal to 1, a score of 2 points is equal to 0.67 and a score of 1 point is equal to 0.33. In the next step, a paired t-test could then be conducted to compare the difficulty level of changed items of CFT 20 and CFT 20-R.

Results

Test versions and Form A or B of the extension were randomly distributed between participants (76 or 48% of participants received test version 1 of the CFT 20/-R, 78 or 49.7% received form A of the extension). Test version 1 (TV1) refers to the conducted version of utilising CFT 20-R as base and has the not-overlapping items of CFT 20 at the end. The other way around applies to test version 2 (TV2). For better understanding, TV1 was termed “TV-new” and TV2 was termed “TV-old” when displaying the results below. Table 1 provides descriptive statistics of the transformed IQ scores of parts 1 and 2 of figural intelligence for 1977 and 2015 norms for the two test versions.

For the extension, Form A and B were also inspected separately, but no relevant differences were found: Raw scores for WS were 13.10 (SD = 13.432) and 13.35 (SD = 13.530) and for ZF raw scores of 8.72 (SD = 9.209) and 8.48 (SD = 8.860) were found for Form A and B respectively. It is possible to compare raw values for the extension, since the same number of items were provided for the pseudo parallel test forms A and B.

Table 1

IQ scores of figural intelligence (CFT 20/-R) compared between the test versions

IQ Score CFT	TV-Old (*N* = 81)				TV-New (*N* = 76)			
	M	SD	Min.	Max.	M	SD	Min.	Max.
Part 1 T1	103.64	15.286	70	141	106.50	16.230	63	136
Part 1 T2	110.57	15.513	73	138	113.32	13.863	60	135
Part 2 T1	104.57	13.974	73	132	106.51	13.508	60	130
Part 2 T2	109.42	11.506	71	127	102.17	13.508	60	130

Note. Min. = minimum; Max. = maximum. TV-Old refers to the test version using CFT 20 as base and presenting non-overlapping items of CFT 20-R at the end. TV-New vice versa.

In a next step, mean IQ scores for each intelligence domain were calculated using norms of 1977 (T1) and 2015 (T2), as well as full-scale IQ, which is computed by averaging the three tested intelligence domains. Findings are presented in Table 2, showing an above average scoring sample with a mean full-scale IQ of around 107 and possible range restriction due to a standard deviation of under 10 for both time points. Results varied for each intelligence domain with figural reasoning scores (CFT) ranging from an average of 105 to close to 109 IQ points and a SD of 12 to nearly 14. Verbal assessment (WS) averaged at an IQ of 107 to 108 and a standard deviation of around 12.5 as well as numerical reasoning (ZF) averaging at close to 106 to 108 IQ points and a standard deviation of 12 to 13. The lowest IQ score of 61 was found for the domain of numerical reasoning in T2 and the highest IQ score was also found in T2 at 138 for figural reasoning.

Table 2

IQ results of participants calculated for old (1977) and new (2015) norms

Intelligence Type	M	SD	Minimum	Maximum
Full-scale IQ T1	107.27	9.15	82.17	126.83
Full-scale IQ T2	107.16	9.91	79.17	129.17
IQ CFT T1	108.90	12.45	72	132
IQ CFT T2	105.27	13.82	68	138
IQ WS T1	107.04	12.47	76	126
IQ WS T2	108.03	12.74	75	124
IQ ZF T1	105.87	12.14	67	123
IQ ZF T2	108.20	13.26	61	126

Note. N = 157. T1 refers to old norms (1977); T2 refers to new norms (2015). CFT refers to CFT 20 and CFT 20-R (figural intelligence); WS refers to extension Wortschatz (verbal abilities); ZF refers to extension Zahlenfolge (numerical abilities).

The results described in Table 2 show that participants of the present sample scored above average, which is further illustrated in Figure 1 (see Appendix) depicting the normal distribution curves of all tested intelligence domains and full-scale IQ in both time points. Especially in verbal and numerical reasoning, a certain test ceiling was apparently reached. For example, 19.1% of people (n = 30) scored all points on the ZF, resulting in an IQ of 123 for T1 and 126 for T2. Close to 20% of people reached an IQ of 123/126, while in an average population distribution, an IQ of over 120 is expected for only around 10% of the population (Weiß, 2019). Similar findings occurred for the WS extension of verbal abilities. A full score of 30 points was achieved by n = 21 (13.4%) and making one mistake (raw score of 29) was reached by 32 people (20.4%). This means that 33.8% of our sample reached a verbal IQ of 126 (raw score of 30) or 117 (raw score of 29) compared with norms of 1977 or a verbal IQ of 124 or 120 compared with norms of 2015.

Even though the present sample shows above averaging abilities, we have nevertheless decided to proceed with the following calculations of ANOVAs since with an N of 157, our sample is expected to be big enough to not be influenced by this distribution. Numerous studies have shown the robustness of effects calculated with an ANOVA with an adequate N (e.g., Blanca et al, 2017).

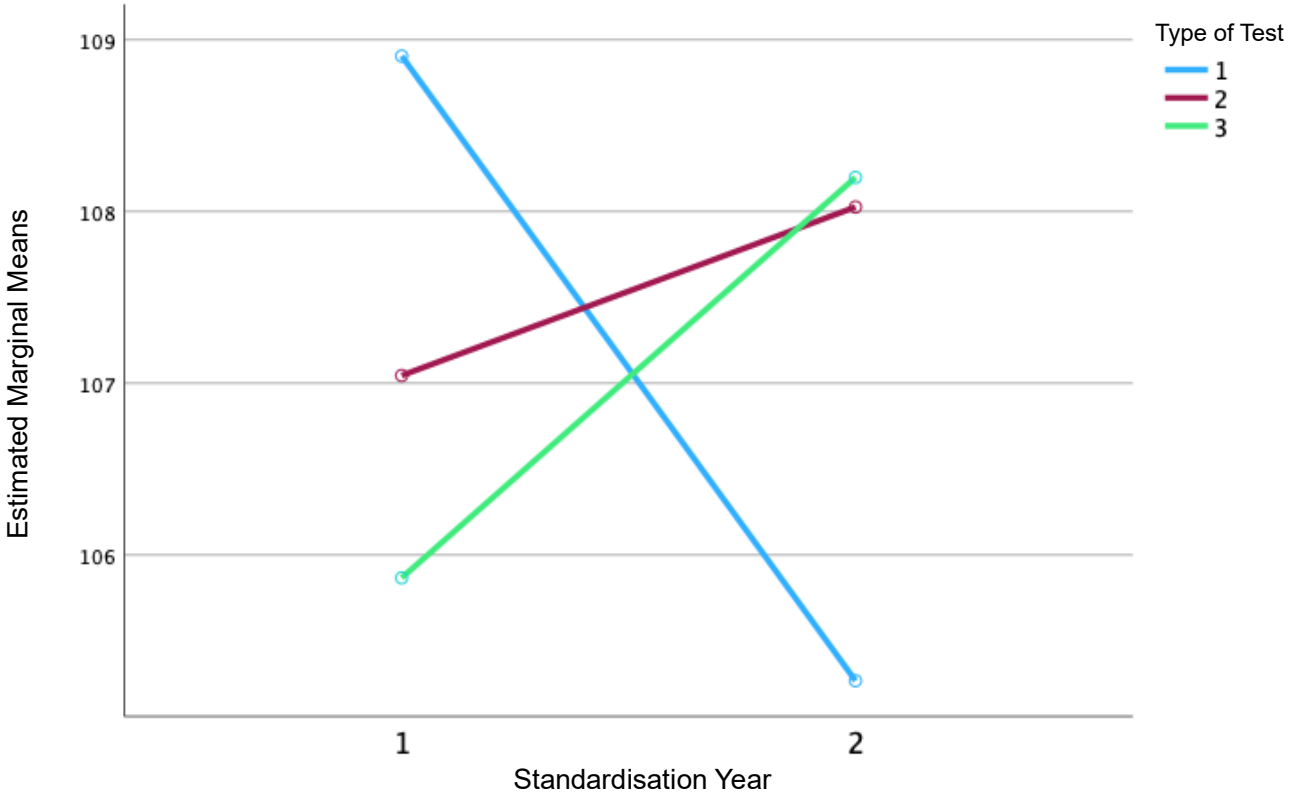
Repeated measure ANOVAs

After transforming raw scores into the respective IQ scores, we calculated two ANOVAs. In the first one, a repeated measure ANOVA was used to compare full-scale IQ with the within-subject factor standardisation year (1977 vs. 2015). Mean full-scale IQs of our sample were 107.27 (T1) and 107.16 (T2) and showed no significant differences (p = .491).

To further examine differences found between the various domain IQs, another ANOVA, a two way repeated measure ANOVA, was performed. Here the three different types of IQ (CFT for figural reasoning, WS for verbal abilities and ZF for numerical reasoning) were compared; standardisation year (1977 vs. 2015) again functioned as one within-subject factor and type of test (CFT vs. WS vs. ZF) functioned as another within-subject factor. Figure 2 shows the respective estimated marginal means. The blue line represents IQ scores of the CFT 20/-R showing a decrease of nearly 4 IQ points between T1 (M = 108.90, SD = 12.45) and T2 (M = 105.27, SD = 13.82). The red line represents verbal IQ scores of WS which increased slightly from T1 M = 107.04 (SD = 12.47) to T2 M = 108.03 (SD = 12.74). The green line, representing the numerical IQ of ZF, also increased, but to a slightly larger extent of approximately 2.5 IQ points from T1 (M = 105.87, SD = 12.14) to T2 (M = 108.20, SD = 13.26).

Figure 2

Estimated marginal means for different types of test across T1 and T2



Note. Types of Test: 1 (blue line) refers to CFT; 2 (red line) refers to WS; 3 (green line) refers to ZF. Standardisation Year: 1 refers to T1 (norms of 1977); 2 refers to T2 (norms of 2015)

A significant interaction was calculated between the factors standardisation year and test type ($p < .001$; *partial* $\eta^2 = .613$). Due to this significant interaction effect, further analysis in form of simple slope analyses was necessary to examine the salience of the respective changes within the intelligence domains. While this slightly deviated from the pre-registration, it was a necessary adjustment since results of an ANOVA may not be interpreted when interaction effects are significant.

Firstly, we compared the three intelligence domains in T1 and T2 using another ANOVA and found significant effects for all three intelligence domains: CFT 20/-R showed a $p < .001$ with a *partial* η^2 of .330; WS a $p < .001$, *partial* $\eta^2 = .237$ and ZF a $p < .001$, *partial* $\eta^2 = .726$. After Cohen (1965) a *partial* η^2 of .01, .06 and .14 is equal to a small, medium and large effect. The reported effect sizes are thus very large but have to

be interpreted in relation to the time span of nearly 40 years between the standardisation years of 1977 and 2015.

A second simple slope analysis was conducted for the three types of intelligence tests. On that account another ANOVA was performed, grouping the intelligence tests by T1 and T2, i.e. comparing 1977 normed results of CFT 20/-R, WS and ZF with each other and doing the same for 2015 normed results. In both time points, significant results with small effect sizes could be reported: T1: $p = .018$; *partial* $\eta^2 = .051$; T2: $p = .026$; *partial* $\eta^2 = .046$.

Item difficulty

In the last step, changed item difficulty levels were tested. Weiß (2019) already reported a difference of complexity for the exchanged items but found no significant differences between tests compared by old and new standardisation. For example, when comparing data from 2015 with that of 1984, he found only slight differences in age degressions for ages 20 to 60 with the two age curves staying nearly identical. Nevertheless, it was important for the present study to test the significance of these item changes and its possible consequences.

Since Part 1 of the CFT 20-R included all items of CFT 20 and items were only added, there was no comparison necessary for Part 1. For Part 2 however, all four subtests included items that were exchanged from T1 to T2 (1 to 3 items per subtest). Paired t-tests were performed and results are provided in Table 4 below. Two-sided *p* values were considered when testing significance in all subtests. All subtests showed significant differences. For Subtest 1, 3 and 4 higher scores were found at T1, indicating that non-overlapping items have become more difficult in T2 than in T1. For Subtest 2, higher results were found at T2, indicating easier items in CFT 20-R compared to CFT 20. Subtest 1 and 2 reported small to medium effect sizes and Subtest 3 and 4 showed large effect sizes.

Table 4

Testing item difficulty by comparing non-overlapping items of CFT 20 and CFT 20-R with a paired t-test

Sub-test	T1		T2		Difference T1 – T2		95% CI		*p*	Cohen's *d*
	M	*SD*	*M*	*SD*	*M*	*SD*	*LL*	*UL*		
1	8.554	1.841	8.172	1.895	.382	1.269	.182	.582	<.001	.301
2	7.441	1.577	7.815	1.793	-.374	.878	-.512	-.236	<.001	.426
3	9.962	2.101	9.026	2.264	.936	1.191	.749	1.124	<.001	.786
4	6.510	1.408	5.633	1.590	.876	.948	.727	1.026	<.001	.737

Note. Two sided *p* values were considered.

Discussion

In the present paper, the more recent version of the Grundintelligenztest Skala 2, the CFT 20, as well as its revision CFT 20-R were examined to test whether a Flynn effect can be found in a current sample in Austria and whether it would differ between intelligence domains. To examine fluid as well as crystallised intelligence, CFT (figural reasoning) as well as the extension Wortschatz (verbal abilities) and Zahlenfolge (numerical reasoning) were used.

Regarding results of full-scale IQ, no significant difference was found between the two time points. However, similar to findings of Lazaridis et al. (2022), vast differences between intelligence domains were found in the present sample. The ambiguous results of positive as well as negative intelligence development for fluid and crystallised intelligence resulted in similar full-scale IQs over time and are therefore discussed further below.

Hypothesis 1 – Figural Intelligence (CFT)

In line with hypothesis 1, the present study found a positive significant difference of approximately 1 IQ point per decade between IQ scores of CFT 20(-R) calculated through norms of 1977 compared with calculations with 2015 norms. However, these differences were not as large as expected, since Flynn (2009) reports gains of around

.30 IQ points per year or 3 IQ points per decade. There are various possible reasons for this occurrence. On one hand, the small figural IQ gains could be due to the reported non-linearity of IQ trajectories and slowing of intelligence gains in developed countries (e.g., Pietschnig & Voracek, 2015), indicating that right now, the Austrian population might be in a phase of lesser intelligence gains. On the other hand, the gains might be underestimated, as a standard deviation of less than 15 suggests a range restriction in the present sample, which can lead to underestimating effects (e.g., Fife et al, 2012). The present sample is also more skilled than the general population with an average of 107 IQ points instead of 100 IQ points for the average population. This is probably due to the primary recruitment of rather well-educated individuals.

Hypothesis 2 – Verbal and Numerical Intelligence (WS & ZF)

In the second hypothesis, we examined whether significant differences could be found for IQ scores of Wortschatz and Zahlenfolge. Testing verbal and numerical abilities, we found that there was in fact a significant difference between IQ scores calculated with 1977 norms compared to 2015 norms, confirming hypothesis 2. However, in contrast with results for figural reasoning, we found a negative development of intelligence. On average, participants scored higher when calculating the IQ with the newer norms compared to the older standardisation. Verbal abilities were found to decrease by approximately 0.25 IQ points per decade and numerical reasoning decreased by approximately 0.50 IQ points per decade. Compared to average gains of 3 IQ points per decade (Flynn, 2009), these losses seem small. These results were nonetheless surprising as for numerical reasoning (ZF) testing fluid intelligence, findings more similar to hypothesis 1 (which also tested a part of fluid intelligence) were expected. Moreover, verbal abilities (WS) were also expected to increase, as a meta analysis by Pietschnig et al. (2010) reported uniform gains of 0.35 IQ points per year for vocabulary (which was also the main focus of WS). However, the latter study analysed data from 1971 to 2007 and now, over 15 years later, the changing trajectory might not be as surprising as it seems at first. As we described above (see Theoretical Background), more and more research from 1995 ongoing has reported a reversal of the Flynn effect (e.g., Flynn & Shayer, 2018). Other studies have also shown a nonlinear development of the Flynn effect in the twentieth century (for an overview see

Introduction or Pietschnig & Voracek, 2015), so it could be possible that right now, Austria is simply in a period of slow development.

In a recent study by Flynn and Shayer (2018) fittingly titled “Does the rot start at the top”, they argue that there might be a decimation of top scorers. This is relevant for the present sample, since, on average, participants scored higher than the common population, putting them into the category of top scorers, for whom Flynn and Shayer (2018) expect to see an intelligence decline first. Moreover, mean age in our sample was close to 30 years, which Flynn and Shayer (2018) report as an age of possible decline. They argue that a decline at one age does not entail a decline at another. Comparing findings to those of the Netherlands, they expect intelligence of preschoolers might stay stable, individuals in school as well as adults might see a decline and older generations might experience gains.

Hypothesis 3 – Differences in Intelligence Domains

The arguments above, however, still fail to explain the very surprising differences reported within fluid intelligence. As described above, fluid intelligence was tested through the CFT 20(-R) (hypothesis 1) and the extension Zahlenfolge (part of hypothesis 2). Due to various reports that fluid intelligence reported stronger IQ gains than crystallised intelligence (e.g., Pietschnig & Voracek, 2015), a similar pattern for the present sample was assumed (hypothesis 3). While it is true that stronger changes could be reported for measures of fluid intelligence compared to crystallised intelligence, those changes went in conflicting directions. This could have various explanations. First, a recent study on German speaking countries by Lazaridis et al. (2022) found differences between intelligence domains to be quite distinct. For example, for crystallised intelligence (comprehension knowledge and general domain-specific knowledge in their intelligence classifications of the CHC model) they reported a positive development. These findings could not be replicated in the present study but are in line with the above mentioned meta analysis of Pietschnig et al. (2010) testing vocabulary. However, it is worth noting that Lazaridis et al. (2022) question their findings around the general domain-specific knowledge as it was operationalised by the English Language Skill Test. Increased English schooling over the past decades in Germanophone countries has most likely influenced the results in this domain.

For fluid intelligence, Lazaridis et al. (2022) reported ambiguous results. For a number sequencing task (which was also applied in the present study for extension Zahlenfolge) they observed small-to-moderate increases and for matrices tests (which was part of the figural testing in CFT in the present study) they found no Flynn effects or only non-trivial ones. This represents the opposite of the present findings where number sequencing tasks eventuated a small decrease and figural reasoning reported substantial intelligence gains. However, while Lazaridis et al. (2022) examined a representative sample of the population, the present sample consisted of above average scoring individuals. Ceiling effects were also present for WS and ZF which could lead to a misestimations of the Flynn effect.

Ceiling effects generally occur when a test or scale is relatively easy, meaning that a substantial percentage of individuals obtain either maximum or near-maximum scores. If this is the case, the true abilities of the high-scoring individuals cannot be measured correctly (Uttl, 2005; Wang et al., 2008).To estimate how severe ceiling effects are, Uttl (2005) found the following: if the group mean performance is within 1 standard deviation of the maximum score of the test, severe ceiling effects can be assumed, thus influencing performance scores of around 25 percent of participants. If the standard deviation is equal to or more than 1.5, then normally less than 10% of examinees scored maximum points and thus performance is unlikely to be limited by severe ceiling effects. As presented above, in the present sample, at T1 participants scored a mean verbal IQ of 107.04 (SD = 12.47) and maximum of 126. At T2, a mean verbal IQ of 108.03 (SD = 12.74) and a maximum of 124 was achieved. A quick calculation shows that the group mean performance is not within 1 standard deviation of the maximum score. However, more than 10% of participants scored maximum points, hinting at ceiling effects for WS.

Similar findings occurred for the numerical reasoning test, at T1 a mean numerical IQ score of 105.87 (*SD* = 12.14) and maximum of 123 (i.e., highest possible raw score of 21) and at T2 a mean numerical IQ score of 108.20 (SD = 13.26) and maximum of 126 was found. Using the above mentioned method to assess ceiling effects, it is shown that the group mean performance is not within 1 standard deviation of the maximum score. However, again the percentage of participants who scored maximum points is well over the recommended 10%, but rather 19.1%, indicating

relevant ceiling effects for ZF nonetheless. It is worth noting that these possible ceiling effects might occur due to the provided age norms. For the CFT 20 standardisation of the extension, ages 13 to 15 were the highest possible ages to compare the present sample with. It seems likely that the present sample, with an average of nearly double this age, has a stronger vocabulary and consequently also achieves high points more easily.

Based on the calculation above, ceiling effects are assumed for the present sample. Therefore, participants possibly could not improve from T1 to T2 as they have already reached a limit. Thus, if no ceiling effects were present, then the small negative development of numerical (and verbal) intelligence might have rather resulted in a stagnation or even positive Flynn effect. Various studies investigating the Flynn effect found that ceiling effects could decrease test score variance, potentially leading to underestimating the Flynn effect (e.g., Sundet et al., 2004; Wicherts et al., 2005). However, these studies mainly report ceiling effects in the second assessment (i.e., the new test or norms). In present findings, ceiling effects seem to be in force for both old and new test versions.

Concluding the arguments above, hypothesis 3 (i.e., fluid intelligence will increase more than crystallised intelligence) was only partly confirmed as results varied within measures of fluid intelligence with a positive Flynn effect for figural reasoning and a small negative trajectory for numerical reasoning, and verbal abilities also declining slightly. However, the development of fluid intelligence showed a steeper development in both cases (figural and numerical IQ) than that of crystallised intelligence. This difference might have even been larger if our sample would have been more representative of the general population. Due to the homogeneity of the sample, effects of range restriction could be present. Moreover, the reported ceiling effects might also bias findings for verbal and numerical abilities. The small negative development could in fact be due to these influences and if this were the case, at least a small positive development might have occurred for WS and ZF in the present data, indicating that the interpretation of present results may have to be ascribed to an artefact instead. A positive development of numerical abilities would also be in line with the small positive Flynn effect of figural reasoning.

Therefore, regarding Weiß' (2019) statement that no evidence can be found for a Flynn effect in the Grundintelligenztest Skala, the present study reports opposing findings.

Limitations

This study is a cross-sectional assessment of the Flynn effect. While this represents an economical method, studies such as Giangrande et al. (2022) caution of using this approach, as they are prone to biases due to possible differences between test versions.

Moreover, as most studies in the literature, the Grundintelligenztest Skala is also based on Classical Test Theory (CTT). As a consequence, test scores are provided, but the latent constructs they are designed to measure cannot be examined. In contrast, tests based on Item Response Theory (IRT) allow the researcher to examine changes in underlying latent abilities. Thus, CTT can illustrate differences in scores, even when no changes in the latent variable are present. Increased scores over time might be due to a "real" intelligence gain or a decrease in the levels of difficulty of test items (Williams, 2013). Missing measurement invariance is not to be treated dismissively (see for example, Beaujean & Osterlind, 2008) and should be considered a limitation of the study.

While the CFT 20 seems to be a widely accepted and reliable measurement of intelligence, recent research has urged scholars to interpret findings of non measurement invariant tests in relation to the Flynn effect with caution (e.g., Gonthier & Grégoire, 2022. In the above mentioned study by Gonthier et al. (2021) for example, the authors found that original findings by Dutton and Lynn (2015) reporting a negative Flynn effect in France, were actually influenced by Differential Item Functioning and showed negative Flynn effects that were not really present in the population. However, other recent studies have found that in some cases, the Flynn effect remains unbiased by whether measurement invariance is present or not (Lazaridis et al., 2022), justifying the current study in continuing its research where no measurement invariance can be achieved for the present study design due to the exchanged items mentioned above.

Moreover, changes in item difficulty between CFT 20 and its revision were found as expected due to reports of Weiß (2019). Except for Subtest 2, all not-overlapping items of the revision showed a higher difficulty level than CFT 20. This would have serious consequences on estimations of the Flynn effect when comparing raw scores.

However, in the present paper, IQ scores were used to compare results. IQ scores were normed according to the respective test items and their difficulty level and therefore do not influence present results of figural intelligence.

For verbal and numerical abilities another limitation arises: ceiling effects. The observed and reported ceiling effects could have made substantial differences to the detected results, possibly leading to underestimation of mean IQ calculations.

Moreover, a missing representation of the population in form of an above average scoring sample should be considered as a further limitation of the study. Descriptive statistics indicated average intelligence scores of approximately 107 IQ points (and maybe even higher when considering ceiling effects in two thirds of the IQ testing) compared to 100 IQ points for the general population and a standard deviation a lot smaller than the expected *SD* = 15. This was likely the result of a sampling bias due to recruitment by word of mouth performed by people with a university level of education. The thereby arising above average scoring sample would in some cases inhibit an ANOVA calculation. However, the present sample size of *N* = 157 should be able to create robust effects nonetheless, even though these effects might be attenuated to a certain extent due to range restriction resulting from undersized standard deviations.

Finally, the structural validity of the newly constructed CFT 20(-R) has to be considered as a limitation. Two new test versions were composed out of CFT 20 and its revision. As we now know, difficulty levels varied significantly between the two tests. One test version started with the easiest items continuously increasing difficulty to finish with the hardest item, while the other test version started with easier (but not easiest) items continued to the hardest items and ending with the easiest items. The Grundintelligenztest Skala is a time sensitive test and might therefore result in individuals not being able to try and solve all the presented items. Therefore, certain item positioning effects might influence results, especially in Part 2 of the second test version.

Contribution

The present paper contributes to research as well as practice in several ways. By identifying significant Flynn effects for the German adoption of the Culture Fair Test, it proposes a possible necessity to update current norms of the CFT 20-R in the near future. Non contemporary norms could have repercussions when the intelligence test is

used for diagnostic reasons or as guidance for important decisions. Moreover, the study provides further evidence for a Flynn effect that varies between intelligence domains in Austria, at least for an above average scoring population. Lastly, the present paper performed an application of the Grundintelligenztest Skala that exceeded figural reasoning, also examining the extension for verbal and numerical abilities, showing possible ceiling effects that should be addressed when revising the CFT 20-R for an adult population. The present study contributes to research with findings of positive as well as negative Flynn effects for Grundintelligenztest Skala 2, contradicting arguments of Weiß (2019) that no Flynn effect is present for the CFT 20 and its revision. However, future research should take the reported ceiling effects for Wortschatz and Zahlenfolge into consideration.

Conclusion

In conclusion, the present paper reports results differentiated according to intelligence domain for the revised Grundintelligenztest Skala 2 (Weiß, 1978; Weiß, 2006). While figural reasoning seemed to improve comparing the sample with norms of 1977 to 2015, results of verbal and numerical abilities both declined to a small extent (with numerical reasoning reporting a slightly bigger effect than verbal abilities). However, these results must be interpreted considering the reported ceiling effects for Wortschatz and Zahlenfolge as participants might not have been able to show improved results from T1 to T2 due to this limit. If this were the case, then the reported small negative development in verbal and numerical intelligence could, in fact, represent a stagnation or even small positive Flynn effect. This would be in line with the small positive development of figural reasoning.

Therefore, present findings hint at small intelligence gains in an above-average scoring population in Austria, opposing Weiß' (2019) statement that no Flynn effect can be found for the Grundintelligenztest Skala 2.

References

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

Blanca, M. J., Alarcón, R., & Arnau, J. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema*, *29*(4), 552–557. https://doi.org/10.7334/psicothema2016.383

Boot, W. R., Kramer, A. F., Simons, D. J., Fabiani, M., & Gratton, G. (2008). The effects of video game playing on attention, memory, and executive control. *Acta Psychologica*, *129*(3), 387–398. https://doi.org/10.1016/j.actpsy.2008.09.005

Brand, C. (1996). *The g factor: General intelligence and its implications*. Wiley.

Cattell, R. B. (1960). *Culture Fait Intelligence Test, Scale 2 (Handbuch)* (3rd Ed.). Institute for Personality and Ability Testing.

Cohen, J. (1965). Some statistical issues in psychological research. In *Handbook of clinical psychology* (pp. 95–121). B.B. Wolman.

Colom, R., & García-López, O. (2003). Secular gains in fluid intelligence: Evidence from the culture-fair intelligence test. *Journal of Biosocial Science*, *35*(1), 33–39. https://doi.org/10.1017/S0021932003000336

Dutton, E., & Lynn, R. (n.d.). A negative Flynn Effect in Finland, 1997-2009. *Intelligence*, *41*, 817–820.

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

Fife, D. A., Mendoza, J. L., & Terry, R. (2012). The Assessment of Reliability Under Range Restriction: A Comparison of α, ω, and Test–Retest Reliability for Dichotomous Data. *Educational and Psychological Measurement*, *72*(5), 862–888. https://doi.org/10.1177/0013164411430225

Flynn, J. R. (2009). *What is intelligence? Beyond the Flynn effect* (Expanded paperback ed.). Cambridge University Press.

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

Foreman, M. M. (1982). Culture-Fair Testing. *The International Schools Journal*, 71–82.

Giangrande, E. J., Beam, C. R., Finkel, D., Davis, D. W., & Turkheimer, E. (2022). Genetically informed, multilevel analysis of the Flynn Effect across four decades and three WISC versions. *Child Development*, *93*(1). https://doi.org/10.1111/cdev.13675

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

Hagmann-von Arx, P., Lemola, S., & Grob, A. (2018). Does IQ = IQ? Comparability of Intelligence Test Scores in Typically Developing Children. *Assessment*, *25*(6), 691–701. https://doi.org/10.1177/1073191116662911

Hoff, E. (2003). The Specificity of Environmental Influence: Socioeconomic Status Affects Early Vocabulary Development Via Maternal Speech. *Child Development*, *74*(5), 1368–1378. https://doi.org/10.1111/1467-8624.00612

Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen. Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells. *Diagnostica*, *28*, 195–226.

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

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *I-S-T 2000 R. Intelligenz-Struktur-Test 2000 R* (2., erweiterte und überarbeitete Aufl.). Hogrefe.

Lynn, R. (2009). What has caused the Flynn effect? Secular increases in the Development Quotients of infants. *Intelligence*, *37*(1), 16–24. https://doi.org/10.1016/j.intell.2008.07.008

Lynn, R. (2011). *Dysgenics: Genetic deterioration in modern populations* (2nd. ed.).

Ulster Institute for Social Research.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. https://doi.org/10.1016/0883-0355(89)90002-5

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

Must, O., & Must, A. (2013). Changes in test-taking patterns over time. *Intelligence*, *41*(6), 780–790. https://doi.org/10.1016/j.intell.2013.04.005

Nyborg, H. (2012). The decay of Western civilization: Double relaxed Darwinian selection. *Personality and Individual Differences*, *53*, 118–125.

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., 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., & 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 Psychologie*, *2*, 267–283.

Rindermann, H. (2008). Relevance of education and intelligence at the national level for the economic welfare of people. *Intelligence*, *36*(2), 127–142. https://doi.org/10.1016/j.intell.2007.02.002

Schroeder, D. (2019, July 11-13). *A negative Flynn effect in recent cognitive ability scores* [Paper presentation]. 20th Annual Conference of the International Society for Intelligence Research, Minneapolis, MN.

Troche, S. J., Wagner, F. L., Schweizer, K., & Rammsayer, T. H. (2016). The Structural Validity of the Culture Fair Test Under Consideration of the Item-Position Effect. *European Journal of Psychological Assessment*, *35*(2), 182–189. https://doi.org/10.1027/1015-5759/a000384

Uttl, B. (2005). Measurement of Individual Differences: Lessons From Memory Assessments Research and Clinical Practice. *American Psychological Society*, *16*(6).

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement Invariance. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01064

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating Ceiling Effects in Longitudinal Data Analysis. *Multivariate Behavioral Research*, *43*(3), 476–496. https://doi.org/10.1080/00273170802285941

Weiß, R. H. (1972). *Grundintelligenztest, Skala 2, CFT 2. Handanweisung für Durchführung, Auswertung und Interpretation*. Westermann.

Weiß, R. H. (1978). *Grundintelligenztest Skala 2 (CFT 20)*. Hogrefe.

Weiß, R. H. (1998). *Grundintelligenztest Skala 2 (CFT 20)* (4., überarbeitete Auflage). Hogrefe.

Weiß, R. H. (2006). *Grundintelligentest Skala 2 - Revision (CFT 20-R)*. Hogrefe.

Weiß, R. H. (2007). *Wortschatztest (WS-R) und Zahlenfolgetest (ZF-R) - Revision. Ergänzungstest zum Grundintelligenztest CFT 20 (Handanweisung)*. Hogrefe.

Weiß, R. H. (2019). *Grundintelligentest Skala 2 - Revision (CFT 20-R)* (2. überarbeitete Auflage mit aktualisierten und erweiterten Normen). Hogrefe.

Weiß, R. H., & Osterland, J. (2013). *CFT 1-R Grundintelligenztest Skala 1 - Revision*. 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

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

List of tables

List of Figures

Appendix

Figure 1a

Full-scale IQ T1

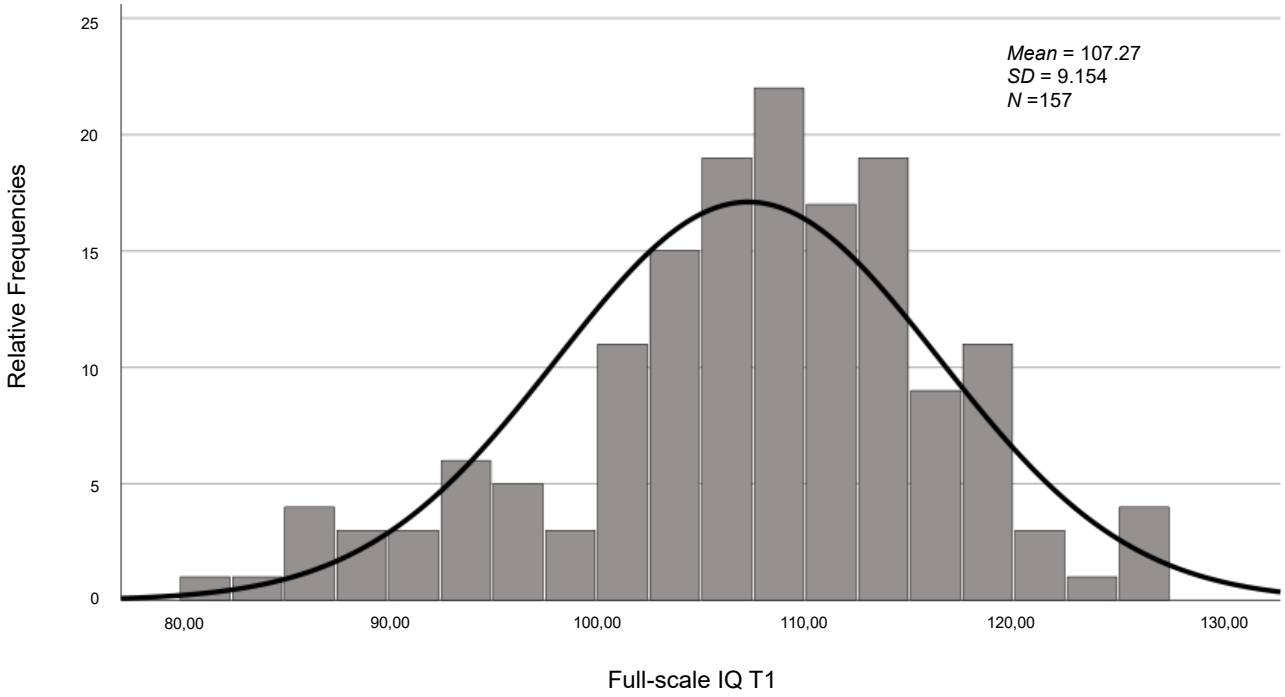


Figure 1b

Full-scale IQ T2

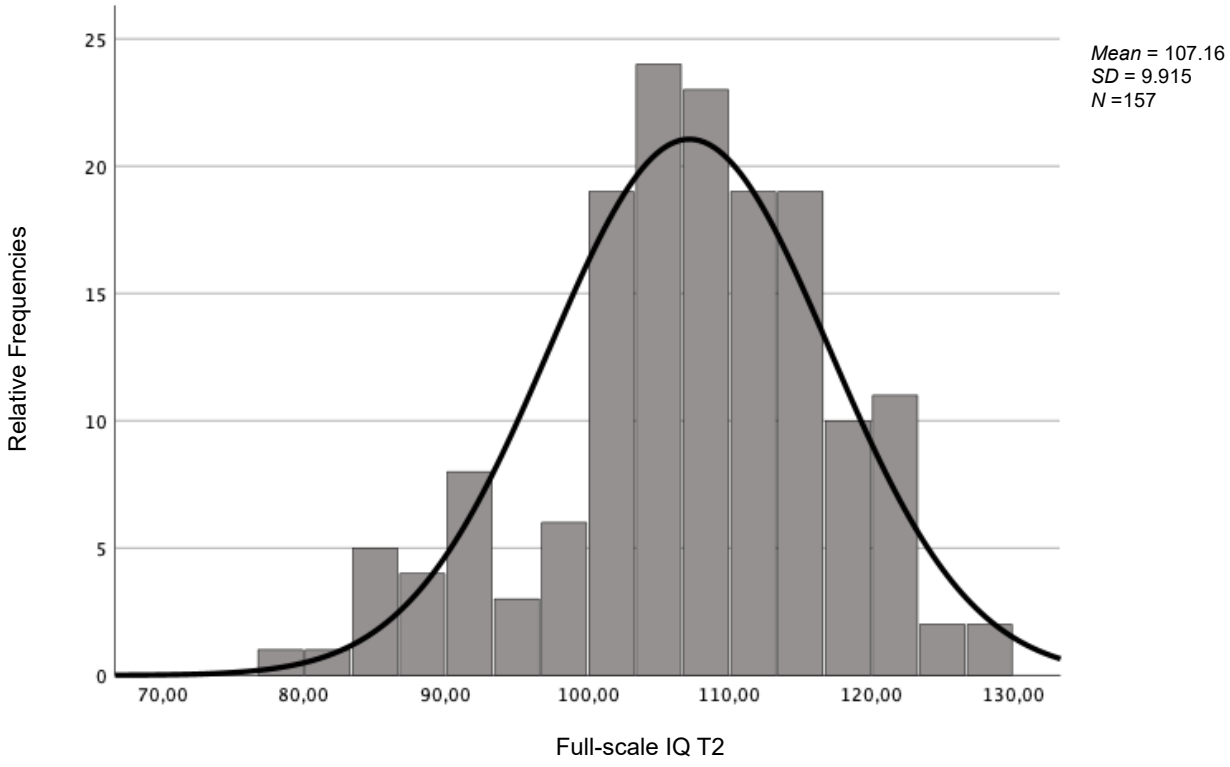


Figure 1c

Figural IQ (CFT) T1

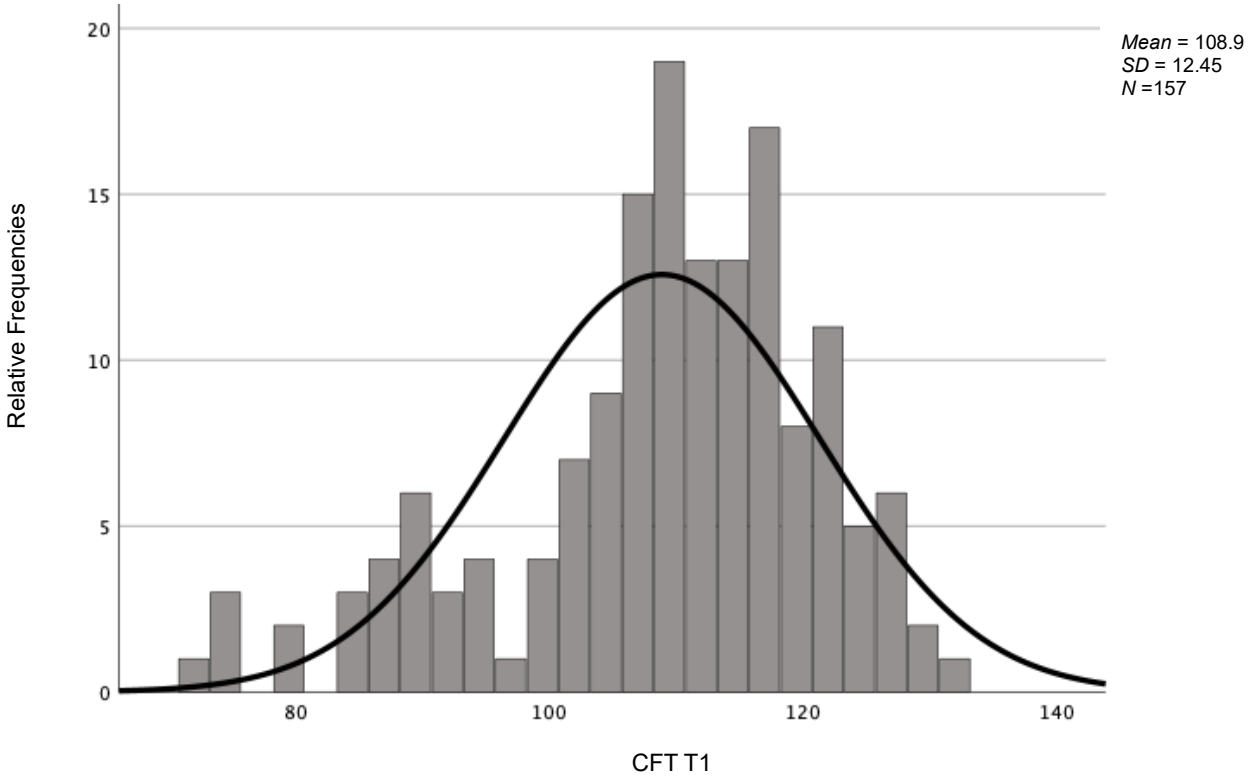


Figure 1d

Figural IQ (CFT) T2

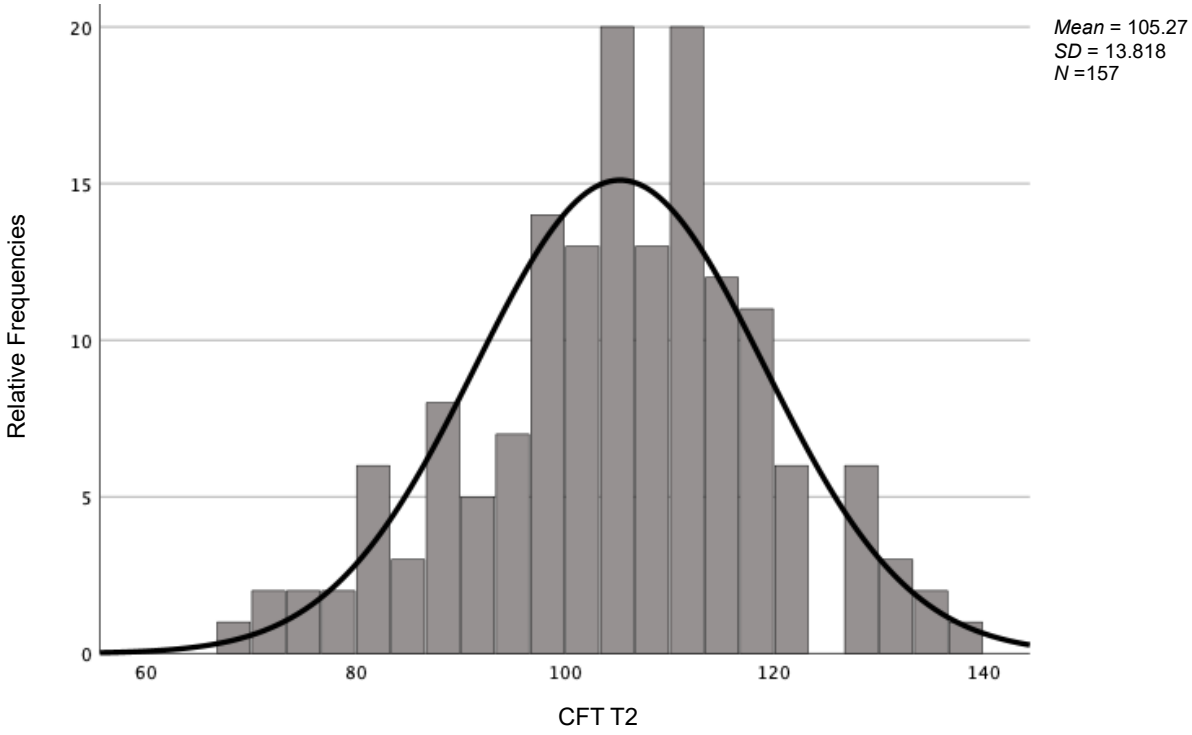


Figure 1e

Verbal IQ (WS) T1

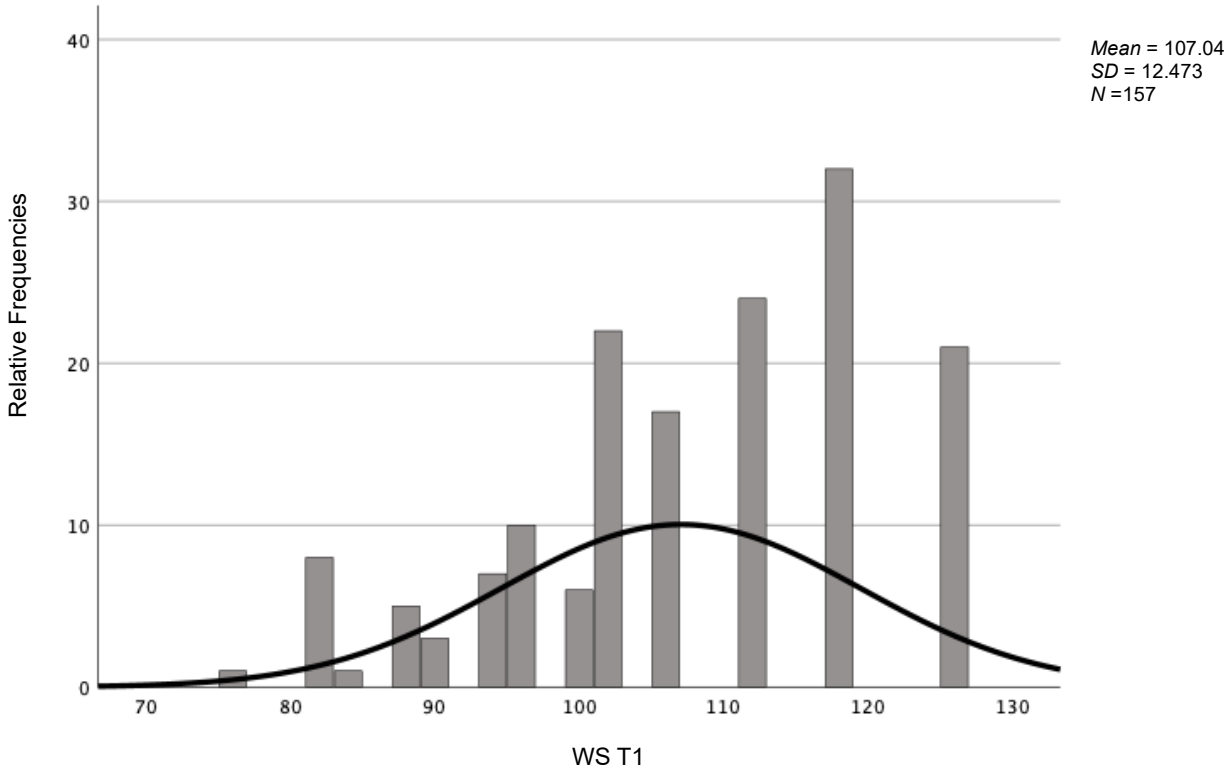


Figure 1f

Verbal IQ (WS) T2

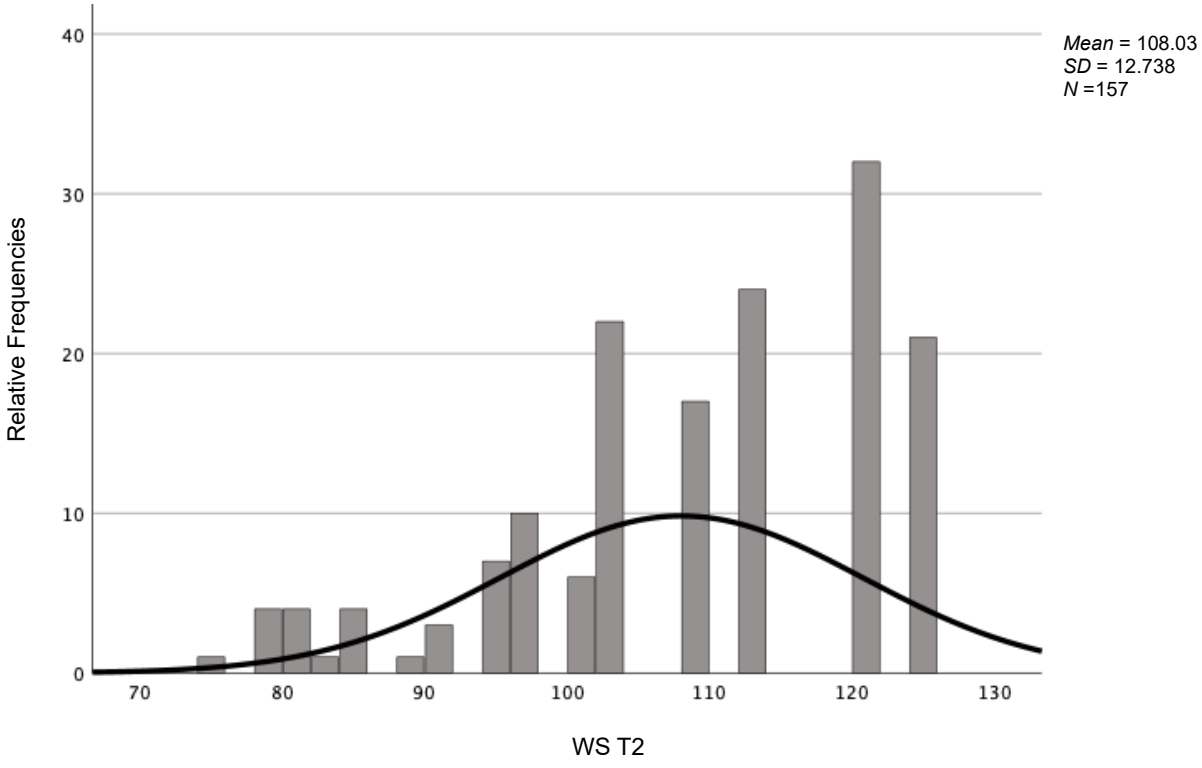


Figure 1g

Numerical IQ (ZF) T1

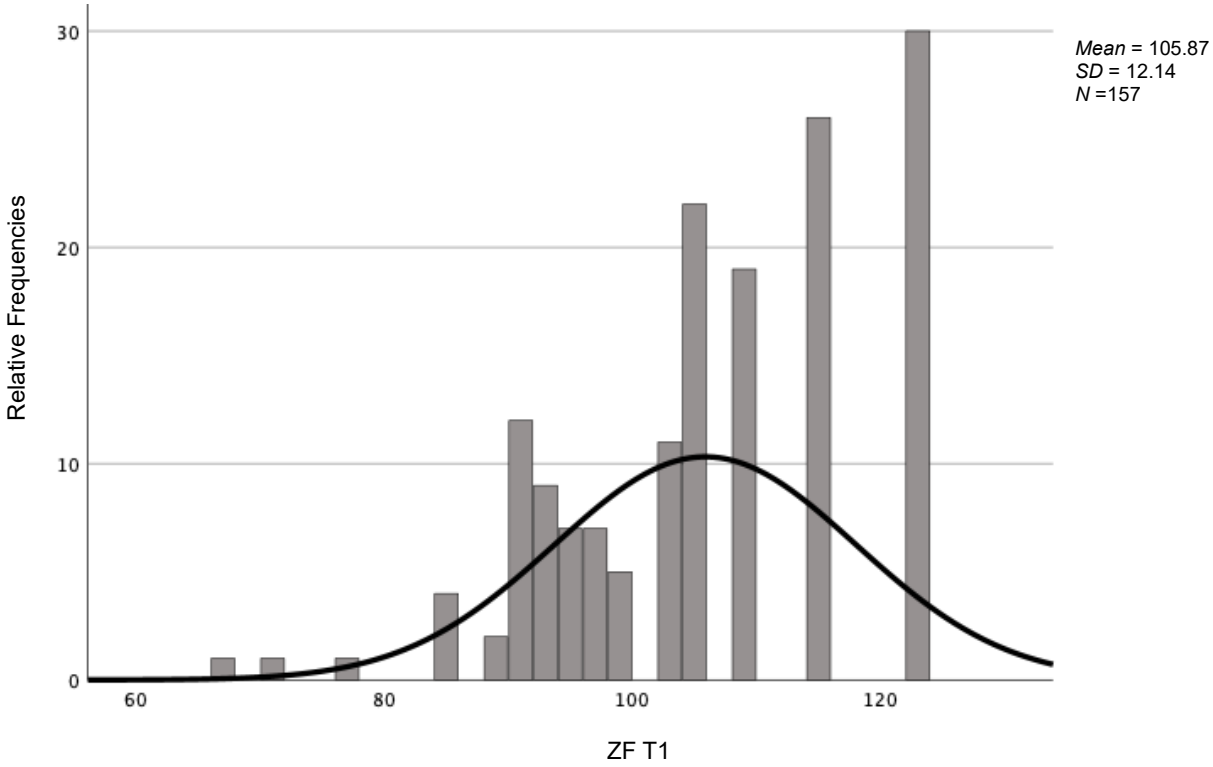
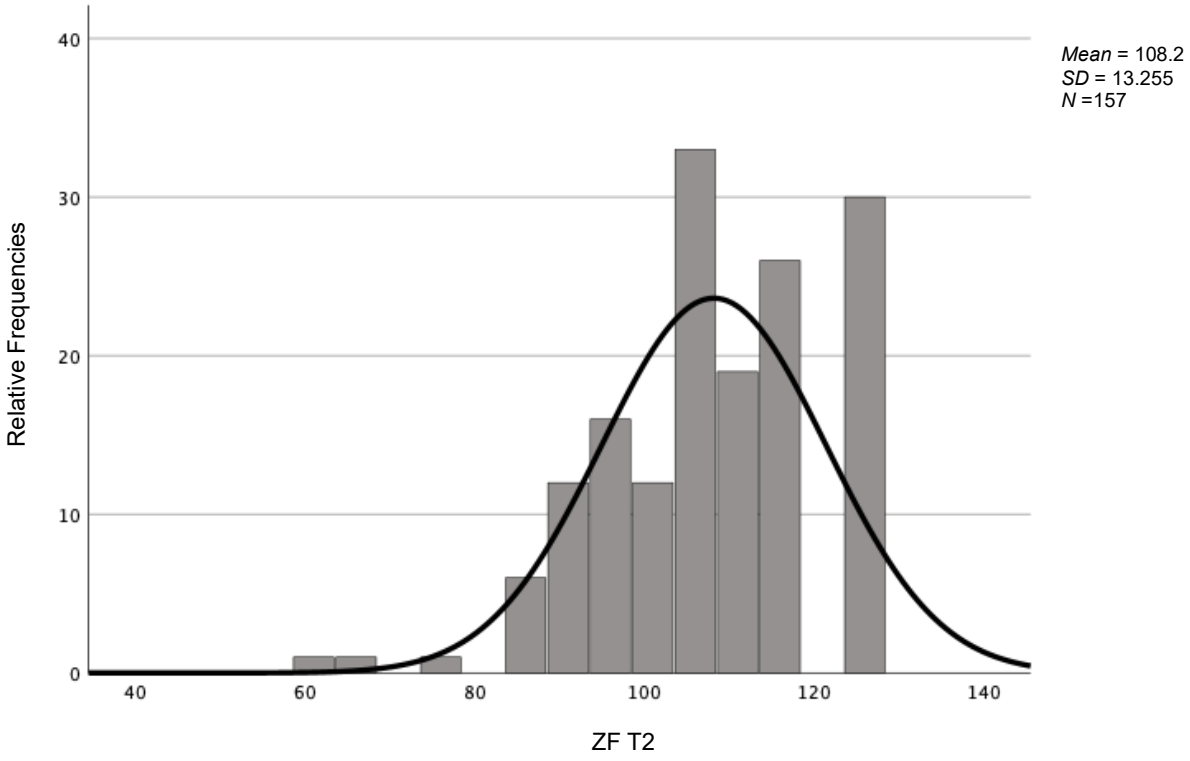


Figure 1h

Numerical IQ (ZF) T2



Abstract

Generational intelligence gains (i.e., the Flynn effect) have been widely observed over the last century and show differences in countries, age and intelligence domains. However, in recent decades, observations of inconsistent patterns in form of stagnations or even reversals of intelligence development have been reported. The present paper investigates the German adoption of the Culture Fair Intelligence Test, the Grundintelligenztest Skala 20 and its revision CFT 20-R for possible Flynn effects on a healthy, German speaking sample (*N* = 157). Using a repeated measure ANOVA, results were differentiated according to intelligence domain when comparing IQ scores to the respective norms of 1977 and 2015. Figural reasoning (CFT) showed intelligence gains, but verbal (extension *Wortschatz*) and numerical abilities (extension *Zahlenfolge*) reported small intelligence losses. However, measurements of verbal and numerical IQ showed influences of ceiling effects. When considering this bias, results may then be interpreted as a stagnation or even small positive Flynn effect, suggesting a similar trend to that of figural intelligence. Therefore, present results of an above-average scoring sample indicate a stagnation or at least slowing of intelligence gains in Austria.

Keywords: Flynn effect, intelligence, Culture Fair Intelligence Test, Grundintelligenztest Skala 2

Zusammenfassung

Intelligenzzuwächse über Generationen (d.h., der Flynn-Effekt) sind im letzten Jahrhundert vielmals beobachtet worden und weisen Unterschiede zwischen Ländern, Altersgruppen und Intelligenzbereichen auf. In den letzten Jahrzehnten wurden jedoch inkonsistente Muster in Form von Stagnationen oder sogar Umkehrungen der Intelligenzentwicklung beobachtet. Die vorliegende Arbeit untersucht die deutsche Adaptation des Culture Fair Intelligence Test, die Grundintelligenztests Skala 20 und dessen Revision CFT 20-R auf mögliche Flynn-Effekte an einer gesunden, deutschsprachigen Stichprobe (N = 157). Unter Verwendung einer Repeated Measure ANOVA wurden IQ-Werte mit den jeweiligen Normen von 1977 und 2015 verglichen und differenzierte Ergebnisse hinsichtlich Intelligenz-Domäne gefunden. Figurales Denken (CFT) zeigte Intelligenzgewinne, aber verbale (Erweiterung Wortschatz) und numerische Fähigkeiten (Erweiterung Zahlenfolge) wiesen geringe Intelligenzverluste auf. Die Messungen des verbalen und numerischen IQ zeigten jedoch Einflüsse von Deckeneffekten. Wenn diese Verzerrung berücksichtigt wird, können die Ergebnisse als Stagnation oder sogar als leicht positiver Flynn-Effekt interpretiert werden, was auf einen ähnlichen Trend wie bei der figuralen Intelligenz hindeutet. Die vorliegenden Ergebnisse einer überdurchschnittlichen Stichprobe deuten also auf eine Stagnation oder zumindest Verlangsamung des Intelligenzzuwachses in Österreich hin.

Schlüsselwörter: Flynn Effekt, Intelligenz, Culture Fair Intelligence Test, Grundintelligenztest Skala 2