MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Domain differences in the Flynn effect: A preregistered cross-sectional assessment of fluid and crystallized intelligence."

verfasst von / submitted by

Vivienne Matev, BSc.

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

Master of Science (MSc)

Wien, 2021 / Vienna 2021

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

1. Introduction

In our society, the use of intelligence testing is widely accepted and common, for instance for the assessment of children with learning impairments, the selection of optimal employees in a workplace setting as well as in the clinical field for a deeper understanding of people's cognitive functioning. The broad range of uses for IQ tests are still expanding and nowadays many fields depend on the assessment of individual's IQ's.

The possibility of measuring the intelligence level of individuals has its roots in early intelligence theories, as for example in the work of Charles Spearman (1927), who postulated a "Two-Factor-Model" of intelligence. According to that theory, two main factors of intelligence can be measured in intelligence tests and therefore describe individual's intelligence levels. Identified through a factorial analysis, the two factors consist of the *g*-factor, which is a general factor of intelligence, that describes the one ability all cognitive tests measure and have in common; and the second factor, the *s*-factor, describing specific abilities and the uniqueness of each cognitive skill. These specific abilities vary within as well as between individuals and consist of a broad range of different acquired skills (Jensen, 1998).

To examine this theory, several different intelligence tests have been correlated in order to determine the *g*-factor, hence creating a bigger picture of what intelligence actually is and consists of. However, even though the two factors postulated by Spearman in 1927 do seem to exist and make up for parts of the human intelligence, additional factors apart from the general factor of intelligence are needed in order to explain the construct of intelligence and how it can be tested (Jensen, 1998). Further analysis of not only the general intelligence but especially the specific abilities shape the knowledge of human intelligence as we know it today.

In this regard, Thurstone (1938) examined so called primary mental abilities based on findings from a factor analysis in his book "The Vectors of the Mind" (Thurstone, 1935). Contrary to the beliefs of Spearman (1927), he did not support the hypothesis of one general factor of intelligence but rather stands for the notion of several group-factors that contribute to the human intelligence as a whole. He postulated seven primary mental abilities of intelligence that consist of verbal comprehension (V), word fluency (W), number (N), spatial ability (S), associative memory (M), perceptual speed (P), and reasoning (R). These group factors represent clusters of correlations between certain types of test items and subtest (Thurstone, 1938). However, since most of Thurstone's primary mental abilities are

correlated, it can be argued, that a general factor of intelligence underlies the individual factors.

In addition to Spearman's "Two-Factor-Theory" (1927) and the primary mental abilities model postulated by Thurstone (1938), other intelligence theories have been evaluated, one of the most influential for today's intelligence testing as well as diagnostics being the Cattell-Horn-Carroll model of intelligence (CHC-model; McGrew, 1997, 2005, 2009). The CHC-model is a combination of two prior intelligence theories, Raymond B. Cattell's and John L. Horn's extended "Gf-Gc-theory" of intelligence (Horn, 1991), which amongst fluid and crystallized intelligence proposes seven further cognitive ability clusters, and John B. Carroll's "Three-stratum-theory" (1993), which also supports the notion of several distinct individual differences in cognitive abilities and classifies them into three stratums. Based on these two models, the CHC-model was derived by carrying out a factor analysis of different intelligence test items and several re-analyses of existing studies using factor analysis as a main method. The Cattell-Horn-Carroll model (McGrew, 1997, 2005, 2009) consists of three hierarchically structured stratums, as proposed in the "Three-stratum-theory" by Carroll (1993); stratum III, which entails the *g*-factor of intelligence and therefore describes the general intelligence of individuals, stratum II, which holds ten broad abilities of human intelligence and stratum I, which is split up into 73 narrow abilities. Stratum II of the CHC-model is made up of cognitive abilities on the one hand, and academic skills on the other hand, overall consisting of Cattell and Horn's distinct intelligences ("Gf-Gc-theory"; Horn, 1991) fluid intelligence ("Gf"), crystallized intelligence ("Gc"), short term memory ("Gsm"), visual processing ("Gv"), auditive processing ("Ga"), long term memory ("Glr"), processing speed ("Gs"), reaction speed ("Gt"), quantitative knowledge ("Gq"), and reading- and writing skills ("Grw"). To this day, the Cattell-Horn-Carroll model of intelligence is one of the most influential theories in the field of intelligence and has led to several intelligence test batteries that build upon the model postulated in this theory.

Making use of two Hamburg Wechsler Intelligence Scales (Hamburg Wechsler Intelligence Test for Adults; Hardesty & Lauber, 1956; Hamburg Wechsler Intelligence Test for Adults Revised; Tewes, 1991) in the current study, which amongst other intelligence tests are also broadly based on the "Gf-Gc-theory" (Horn, 1991) and the CHC-model of intelligence (McGrew, 1997, 2005, 2009), it is important to understand the concept of these models. Specifically, Cattell and Horn's evaluation of fluid and crystallized intelligence ("Gf"; "Gc"; Horn, 1991) and their distinct roles in the assessment of the Flynn effect, which describes observations of rising intelligence test performance in the general population over

the last decades (Flynn, 1984; Pietschnig & Voracek, 2015), is the main focus of this thesis. The broad ability domain "fluid reasoning" ("Gf"), in this study also named "fluid intelligence", describes the ability to solve abstract problems and mental operations, such as drawing inferences, identifying relations, comprehending implications and reasoning (McGrew, 2009). In intelligence tests, items that mainly test fluid intelligence usually do not get changed as much during revisions due to the fact that fluid reasoning does not rely on prior knowledge, thus items do not tend to get obsolete. For instance, completing a puzzle is a mental operation that does not rely on cultural or prior knowledge, thus individuals are able to solve an item based on that mental operation solely relying on their fluid intelligence capacity. In contrast, crystallized intelligence, in CHC-theory referred to as "comprehension knowledge" ("Gc"), describing mostly verbal or language-based knowledge, that is acquired through education and acculturation (McGrew, 2009), has to be changed more often in intelligence test revisions due to changing cultural circumstances and a broadening of generation- and culture specific knowledge. An example would be a question about the current president, which can only be answered correctly if that president is still in office at the time the question is asked. Otherwise, this item would become obsolete and individuals would have a harder time solving it, despite a possible high crystallized intelligence ability. Therefore, items testing crystallized intelligence in IQ tests tend to get obsolete and have to be adapted in test revisions. In regards to the Flynn effect, the present study specifically focusses on differences in fluid and crystallized test score changes over time and potential explanations for domain differences in the generational IQ changes.

1.1 From intelligence theories to intelligence scales

Focusing on intelligence testing today, one of the most popular intelligence scales used is the Hamburg Wechsler Intelligence Test, two versions of which are used in this study, introduced in 1939 with the publication of the Wechsler-Bellevue Intelligence Scale (Boake, 2002). The most recent Wechsler Scale (Wechsler Adult Intelligence Scale – Fourth Edition; Petermann, 2012) includes ten subscales as well as five additional ones for optional use, most of which can be traced back to the very beginning of intelligence testing. The subtests are roughly categorized into five sections, namely verbal comprehension, which consists of the subscales Similarities, Vocabulary, Information, and the optional subtest Comprehension; perceptual reasoning, including the subtests Block Design, Matrix Reasoning, Visual Puzzles, Picture Completion, and Figure Weights, the last two of those being optional tests as well; working memory, consisting of the subtests Digit Span, Arithmetic and the optional test Letter-

Number-Sequencing; and lastly processing speed, which includes the subtests Symbol Search, Coding, and the optional test Cancellation. Tasks akin to the digit span test have been already used by Francis Galton in 1887 to identify school children's mental capacity, measuring abilities in working memory, attention, encoding and auditory processing through listening to sequences of numbers orally and repeating them as heard, in reverse order, and in ascending order (Boake, 2002). Additionally, almost all of the subtests have been derived from the first standardized intelligence test by Alfred Binet and Theodore Simon (1905), which has been developed to identify learning-impaired school children in Paris. Included were tests to assess language skills, memory, reasoning, digit span as well as psychophysical judgements (Binet & Simon, 1905).

In 1908, Binet and Simon revised their intelligence scale by grouping tests into age levels ("year scale") and quantifying intelligence into intellectual levels, which was later changed to the so-called mental age. Even though it was criticized, that their test was overemphasizing verbal skills and therefore excluding people with lower vocabulary or foreigners, the test became widely used and was soon translated into the English language, which led to an even further distribution of use in America (Boake, 2002).

Another revision took place shortly before World War I, by the American psychologists Robert Yerkes and James Bridges, who changed the year scale system into a point scale, that is now still used in the Wechsler Scales (Thorndike & Lohmann, 1990). The point scales system basically groups items that are similar into subtests, which are then administered by starting with the easiest test item and proceeding in order of item difficulty.

The most important change to the Binet-Simon intelligence test was introduced by Lewis Terman, who replaced the mental age with the now well-known intelligence quotient (IQ), nowadays used in many psychological areas, such as for instance in educational settings as well as diagnostics. Additionally, he extended the scale into adulthood and added arithmetic reasoning items and a form board. This version of the Test – the Stanford-Binet Intelligence Scale (Terman, 1916) – should then become the main intelligence measure in America (Boake, 2002).

Since the Stanford-Binet Intelligence Scale was based on mostly verbal subscales, more practical tests were introduced for people with limited English-language skills or hearing-impairments, such as the Pictorial Completion test, where missing pieces of a picture have to be selected in order to test the ability to quickly perceive visual details (Healy, 1914, 1921). These non-verbal tests, also termed "performance tests" were later combined into the Pintner-Paterson Performance Scale, an intelligence test battery for hearing-impaired school

children (Pintner & Paterson, 1917). Although the scale does not exist anymore today, some items were preserved and are used in the Wechsler Intelligence Scales to this day, such as, for example, items of the Object Assembly subtest (Boake, 2002).

After Alfred Binet developed the first standardized intelligence tests to identify learning-impaired children in 1900, the tests quickly turned into a common measure for the U.S. military to place new recruits in different positions according to their unique skills and abilities. These were the main building blocks for the Wechsler Intelligence Scales, as for instance the Arithmetical Problems subtest, the Information subtest and the Practical Judgement subtest were sourced from the group military examinations and adopted in the Wechsler Scales. Additionally, performance subscales, that were introduced in the Beta tests of the group examinations, were sourced, as for example the Digit Symbol test (Yerkes, 1921), parts of the Picture Completion test (Yerkes, 1921) as well as some items from the Picture Arrangement test (Yerkes, 1921).

In addition to the group examinations, the military intelligence testing also entailed individual examinations, that served as a major source for the Wechsler Intelligence Scales, particularly the subtests Arithmetic Reasoning, Likenesses and Differences, and the Comprehension test (Boake, 2002).

Wechsler conducted several of the individual military examinations, where the first ideas of the later established Wechsler Scales developed. As a result of the observation, that in the military examinations several separate tests were administered, such as the Stanford-Binet Scale as well as performance tests, Wechsler came to the conclusion that a combination of verbal and non-verbal subtests in one single intelligence test would be advantageous for the future of intelligence testing. Furthermore, he advocated for a renewal of the statistical approach of calculating the IQ, since the Stanford-Binet test was more suitable for children than for adults (Wechsler, 1981).

Following the development of the first Wechsler Scale, the Wechsler-Bellevue Intelligence Scale, and the further use of intelligence testing in general, a lot of criticism arose, concerning standardized means to broadly test the intelligence of children and adults in educational environments. Most of the disapproval addressed the unfair stratification of test-takers by race, sex, class, and culture, as well as the ignorance of creativity, emotional intelligence and practical know-how (Benson, 2003). Through the introduction of regular revisions of the existing standardized test measures, such as the Wechsler Intelligence Scale for children as well as for adults and the Stanford-Binet Intelligence Scale, such gaps in generalizing the test results of these intelligence measures were hoped to be accounted for.

Although over time alternative intelligence theories have been proposed that are slowly being implemented into intelligence testing, the main methods used for the assessment of people's IQ are still based on the models proposed by Binet and Wechsler, who significantly shaped the now dominating use of the Intelligence Quotient (IQ).

1.2 The Flynn effect

When analyzing intelligence test performance over the years, an incline in performance in the general population was observed and first noticed as a cohort effect by Schaie and Strother in 1968, which is now formally termed the Flynn effect. This effect can be generally noticed in the fluid and crystallized domains of intelligence, even though its intensity varies throughout the domains. In intelligence tests, fluid IQ can be assessed through reasoning-based tasks that can be solved with no prior knowledge and are usually non-verbal, performance scale based, whereas crystallized IQ is assessed through knowledge-based tasks that cannot be solved with reasoning and are mainly verbal subtests (Pietschnig & Voracek, 2015).

The Flynn effect, which is a now well-known phenomenon in the field of intelligence, describes the rising levels of intelligence test scores, that have been observed in the American population as well as other economically developed and less-developed countries through the preceding decades (Flynn, 1987). In their book "The Bell Curve" (1996), Herrnstein and Murray firstly named the effect after Flynn, who researched the effect thoroughly. However, the observation of increasing IQ scores was present way before that. For instance, a study conducted by Runquist in 1936 indicated an incline in IQ from the years 1929 to 1932, which was supported by several further studies in the following years. In conclusion, after the year 1973 more than ten independent studies were published, showing generational IQ gains with a mean of 3.65 IQ points per decade. Nevertheless, in these early years the observed effects of increasing intelligence test performance were attributed to sampling errors (Runquist, 1936). This was the case until 1968, when Schaie (Schaie & Strother, 1968) first identified the IQ gains as a cohort effect before Flynn (1984) started researching the phenomenon more intensely, specifically regarding test norm changes. He argued, that in order to calculate IQ scores, specific norms for the transformation of raw scores to standardized IQ scores have to be established. Therefore, means and standard deviations of a representative norm sample have to be calculated, which can then be used to standardize scores. When a revision of test versions takes place, the norms of each new test rely on those from the original test version in order to maintain validity. Hence, generational IQ gains can be observed through analyzing the norm populations raw scores for each test version, where higher raw scores on the newly

normed tests would be an indicator for such gains. In turn, individuals should be able to perform better on tests with older test norms, since the norm populations mean IQ's were lower than in the revised test versions. Following this reasoning, this would only be the case if the tests content has stayed the same regarding its difficulty and structure and only the norms have been adapted. If no generational IQ changes were present, two representative norm samples should achieve the same raw scores on two equivalent tests taken years apart. Yet, if the second norm sample scored five points higher than the first one, this would indicate an increase of five points in the general population over that time period. When using a cross-sectional assessment to investigate this phenomenon, two test versions of the same intelligence test – the original as well as the revised version – would be administered to one sample. As Flynn observed, almost all individuals scored higher on the original test version compared to the revised one, suggesting a positive Flynn effect due to the norm populations IQ increases (Flynn, 1984, 1987).

The impact of these findings regards many subjects in the field of psychology, for instance developmental psychology, education, organizational psychology, clinical psychology and diagnostics, as for example in the diagnosis of intellectual disabilities. A false assessment of intelligence when using older test norms, specifically an over-estimation of about 0.3 IQ points per year, might influence the treatment and support a child with mental disabilities receives. Oftentimes, an actual intellectual disability is therefore mistaken for a learning impairment, hence leading to false interventions for the individuals. However, also judicial decisions might be affected by the Flynn effect, as Flynn (2006) points out. Hence, in legal matters, IQ values can influence the sentencing of an individual, as for instance in certain US-states no death penalties are executed when intellectual disabilities and an IQ under 70 are identified. A false IQ score can therefore have detrimental consequences (Flynn, 2006).

Interestingly though, the test score changes observed in the Flynn effect vary across countries as well as domains (Pietschnig & Voracek, 2011, 2015; Lynn, 2009; Flynn, 1984), yet the mean IQ gains can be generally located between 3 and 5 IQ points per decade until the year 1998. After that, a small decline in the Flynn Effect could be observed in Denmark, yet globally the declines could be noticed even earlier (Teasdale & Owen, 2008). Consequently, a study by Teasdale and Owen (2005) has examined whether in Denmark the Flynn Effect has potentially ended by assessing intelligence test scores of 549,149 male subjects between 18 and 19 years of age between the years 1959 and 2004 using the Borge Priens Prove (BBP, Rasch, 1980). Results indicated that after the year 1998 intelligence test results had indeed

fallen, compared to the decades before that, showing that the mean scores of the intelligence test used in that study in the years 2003 and 2004 corresponded to those of the year 1990 (Teasdale & Owen, 2005). An interesting finding in that study also revealed that the decline of IQ points after the year 1998 was also accompanied by similar changes in educational levels, meaning that when test scores peaked in 1998, the attendance of advanced-level colleges has also peaked and subsequently fallen the years after. However, the described study has several limitations, including only testing male subjects in Denmark, which could pose some problems in generalizing those findings.

Noteworthy conclusions were also reported by Pietschnig and Voracek (2015) who conducted the first formal meta-analysis on the generational IQ gains that are now known as the Flynn effect, with a total of almost 4 million participants from 31 countries. As past research has shown, the strongest IQ gains were observed in Austria, France, Germany, Israel, Japan, Kenya, the Netherlands, and Spain, with stronger gains in fluid intelligence, as well as larger gains between the world wars. Key findings in the meta-analysis conducted by Pietschnig and Voracek (2015) included the observation of substantially stronger gains in IQ points for fluid intelligence than for crystallized intelligence, numerically 4.1 points of gain in fluid versus 2.1 IQ points in crystallized intelligence per decade. These domain differences were also shown in studies by Lynn (2009) and Flynn (1984). The stronger increases in non-verbal subtests have been already seen in several other studies as well, as for instance a study by Wheeler (1942) brought to light, where gains of 6 IQ points were found in fluid intelligence, compared to an increase of only 2.6 IQ points in crystallized intelligence in a time span of ten years.

Additionally, a negative correlation with the g-factor of intelligence could be observed when analyzing the Flynn effect, implying that specific cognitive abilities are affected by the IQ gains rather than the general intelligence (Pietschnig & Voracek, 2015). Lynn (2009) also found, that IQ gains were greater for children on a low ability level, specifically at the 5th percentile of the IQ distribution (IQ 75) and declining from that point on. Additionally, it was shown that there were no sex differences in the Flynn effect, even though men show a slightly higher test performance in certain domains, suggesting that the observed IQ gains are sex-independent (Pietschnig et al., 2011).

Furthermore, as Teasdale and Owen (2005) already examined, decreasing gains in more recent decades were found, especially in Scandinavian countries, where even a reversal of the IQ gains might be the case. This also supports the finding, that IQ test score gains have not been linear over the decades, but rather nonlinear, as the alternately accelerating and

decelerating gains can show (Pietschnig & Voracek, 2015). Ultimately, there were stronger IQ gains found for adults than for children with large effects for fluid and spatial IQ (IQ score derived from items consisting of mentally rotating an item in order to solve that item), as well as a positive association of IQ gains and gross domestic product per capita (GDP), especially with full-scale (IQ score derived from several subtests testing different intelligence domains), crystallized and spatial IQ (Pietschnig & Voracek, 2015).

1.3 Possible explanations for the Flynn effect

There are several theories on why the Flynn effect occurs, yet most of them cannot explain the domain-specificity observed in the IQ gains. In general, the causes of IQ test score gains can be divided into environmental, biological and hybrid causes, the latter being a combination of both, biological and environmental factors (Pietschnig & Voracek, 2015).

1.3.1 Education

Improvements in the educational system of industrialized societies may play a distinct role in IQ test performance and the rising IQ scores noticed, as Ceci and Williams (1997) amongst others have suggested. This hypothesis is based on the findings of studies published in the early twentieth century, where education and the development of intelligence were significantly correlated (Tuddenham, 1948). Similarly, negative correlations between age and IQ in children who did not attend school regularly could be found, thus leading to a decrease of IQ with increasing age of those children. Specifically, when comparing children who did receive education regularly with those who failed to attend school on a regular basis, a decrease of intellectual abilities of about 5 IQ points per year was noticed (Ceci & Williams, 1997).

Nevertheless, nonsignificant gains and even declining performance on intelligence tests during times when a general IQ incline was observed counter the hypothesis of education as a main cause for the Flynn effect. Additionally, the rising intelligence test performance of the general population affects the fluid domain of intelligence more than crystallized intelligence, yet educational effects would rather point toward crystallized IQ changes (Pietschnig & Voracek, 2015). Still, it has to be said that better schooling and educational systems are important environmental factors when analyzing the Flynn effect, even though when controlling for education, IQ gains have reportedly remained substantial (Pietschnig et al., 2013). According to those findings, the conclusion can be drawn, that education cannot explain the Flynn effect fully, especially not in the domain of fluid

intelligence. Nevertheless, valuable insight as to where parts of crystallized intelligence gains in IQ tests may stem from can be obtained (Pietschnig & Voracek, 2015).

1.3.2 Nutrition

A hybrid factor that could potentially explain intelligence test performance gains over the past decades is nutrition. As Lynn (2009) states, improvements in nutrition in well developed countries have emerged over the course of the twentieth century, leading to an increased mean height and head circumference of well-nourished individuals. It has been shown by several studies, that nutrition has an effect on intelligence test performance as well, especially on fluid intelligence subtests. For instance, Hunt (2011, p. 260) found, that poor nutrition leads to lower IQ test performance, which in turn could explain higher test performance through improved nutrition. This has been observed to a greater degree in fluid intelligence compared to crystallized intelligence, indicating that poor nutrition interferes more with fluid abilities. Moreover, studies have shown, that the use of nutritional supplements for children leads to increases in fluid intelligence and fluid subtest performance in intelligence tests (Lynn, 2009; Pietschnig & Voracek, 2015).

Lynn (2009b) has also suggested nutrition as a factor that could potentially explain the Flynn effect due to improved pre- and postnatal nutrition, which has been shown to help cognitive development and in turn might influence the IQ test performance of these individuals. Nonetheless, because of the fact that higher IQ gains were observed in adults than in children, Pietschnig and Voracek (2015) suggest, that there may be other factors involved that contribute to the rising intelligence test performance.

1.3.3 Pathogen stress

A factor that can also only explain portions of the Flynn effect is pathogen stress, describing difficulties of cerebral development in infancy and early childhood due to infectious diseases. Because brain development has a very high energy demand, it is important that all resources are available to help that process. If large parts of those resources are needed to fight illnesses, cerebral development might be impaired, thus leading to weaker cognitive abilities in later life. This finding is reflected by the observation of a negative correlation between the mean intelligence of a country and the prevalence of infectious diseases, meaning that a higher prevalence of infections in a specific country leads to lower mean IQs in that population (Eppig et al., 2010).

The IQ gains associated with the Flynn effect therefore might have emerged because of increased availability of healthcare and health services in general, better sanitary conditions as well as modern medical assistance, which all lower the prevalence of pathogens like

infectious diseases. Despite those factors, explaining IQ gains in the fluid intelligence domain, those gains were mainly observed in adults, which means that other factors, that do not only emerge in childhood, have to be examined.

1.3.4 Social multipliers

A theory that might explain the IQ gains and especially those found in fluid intelligence is the concept of social multipliers, a theory suggested by Dickens and Flynn in 2001. The theory is based on the two contrary opinions, that the increased intelligence test performance observed in the Flynn effect has either emerged solely due to environmental influences or genetical factors, and combining these two approaches of explaining the Flynn effect into one single model. Dickens and Flynn (2001) suggest a reciprocal influence from one factor to the other, meaning that the inherited, genotypical intelligence an individual exhibits and the environment, with which this individual is interacting in daily life, both contribute to the IQ a person develops through life.

According to them, a feedback loop is created by being exposed to slight environmental advantages, which usually would improve individual performance, which then in turn would lead to an even more advantageous environment and so forth. These beneficial environmental influences that hold the ability to give advantages in specific abilities would act as multipliers for those abilities, hence increasing certain skills. On the other hand, also the IQ of other individuals interacting in that same environment would influence the environment itself, which then has an effect on the mean intelligence of the population of that specific environment.

Because it is dependent on societal emphasis of a specific task or ability, Dickens and Flynn (2001) argue, that intelligence test performance gains would be a result of an intense focus on cognitive abilities in our societies, which, according to their social multiplier model, would subsequently enhance the exposure to cognitively stimulating environments. Consequently, even very small environmental changes could have substantial effects in a relatively short time period.

Although this theory seems promising, empirical studies could not directly test this hypothesis, as Pietschnig and Voracek (2015) write in their meta-analysis. To conclude, the social multiplier theory can explain portions of the gains observed in the Flynn effect and is one of the few theories able to explain a domain specific Flynn effect and the observed patterns in IQ gains. Unfortunately though, an empirical confirmation of the proposed explanation is lacking, due to its complex composition and hardly testable hypothesis.

1.3.5 Life history speed

It is very likely, that most of the theories described above are able to explain certain portions of the increased intelligence test performance observed over the years, which makes an interaction of different factors as an explanation for the Flynn effect likely. This is also the approach taken by Woodley (2012), who argues that a combination of better education, reduced family size, better nutrition, and lower pathogen stress form the key reasons for test score changes in intelligence tests. These interacting factors lead to a differentiation of cognitive abilities and are the basis of the life history speed model. In this theory, slow life history individuals are characterized by fewer lifetime sexual partners, fewer offspring as well as later parenthood, compared to fast life history individuals. Those different life history speeds are favored by different environmental conditions that are driven by the factors specified above (e.g., nutrition, family size, etc.), thus enabling slower life history speed for individuals in safer environments, as for example environments with low pathogen stress, adequate nutrition, and so forth, which in turn are provided with the resources necessary for cognitive ability maturation and differentiation. Each factor that reduces the risk of a high mortality due to unpredictable factors like diseases and insufficient food supply therefore promotes a slower life history speed, which leads to a higher possibility of developing certain cognitive abilities.

This might explain rises in IQ scores, especially for fluid intelligence, in individuals with slower life history speed and would also account for erratic patterns of the Flynn effect, as it does not warrant consistency and strength of IQ changes across countries and time. Even though conditions for a slow life history speed might be given in some countries more than in others and therefore an inconsistent pattern of IQ gains would emerge, a general incline in intelligence test performance due to compensatory effects of other given factors would be observable (Pietschnig & Voracek, 2015). Additionally, the data accounting for a negative correlation between an IQ increase and the *g*-factor of intelligence would also support this theory of life history speed, hence a slow life history individual would experience a decrease of the general intelligence but in return an increase of domain specific abilities (Woodley, 2012).

It has to be noted though, that in their meta-analysis, Pietschnig and Voracek (2015) could not confirm fertility as a negative predictor of IQ gains or decreasing IQ gains after decades with large birth cohorts, concluding that decreasing family size cannot be interpreted as a predictor for intelligence test gains.

1.3.6 Other explanations

Several other theories trying to explain the increasing IQ scores have been proposed, yet most of them lack sufficient empirical data and plausibility or no conclusive evidence could be found in subsequent research. Thus, these theories will only be mentioned briefly in this study.

An environmental factor that was thought of playing a role in explaining the Flynn effect is the exposure to technology. This explanation is based on the belief, that visual analytical abilities are being trained when exposed to modern technical appliances, such as computers, television, video games, etc., as suggested by Neisser in 1997. These visual analytical abilities then would in turn facilitate intelligence test taking. However, no conclusive evidence has been found on the account of technology as a valid factor contributing to IQ point gains in fluid intelligence.

Similarly, decreasing family size, which has been linked to increased cognitive task performance (Zajonc & Mullally, 1997) has been rated as an unlikely environmental explanation of rising intelligence test performance, since evidence is scarce (Pietschnig & Voracek, 2015).

Test-taking behavior has been examined as a potential environmental factor contributing to the Flynn effect, especially in fluid intelligence (Brand, 1987a). That is, because especially in recent years, psychometric test instruments have made use of multiple-choice response formats, which in turn facilitate guessing behavior on difficult test items. Such behavior can be usually seen in fluid intelligence subtests, since those are mostly independent of educational backgrounds of test takers and therefore suit a multiple-choice approach best. Nevertheless, as empirical evidence for this theory cannot support all gains observed in the Flynn effect, this explanation should be interpreted with caution.

Amongst others, Jensen (1998, pp. 189-197) found positive associations between average allelic heterozygosity and cognitive task performance. This allelic heterozygosity is increased by the mating of genetically dissimilar individuals and is also referred to as hybrid vigor. As allelic heterozygosity is increasing during the past decades, the explanation of hybrid vigor as a possible cause for the Flynn effect cannot be denied (Mingroni, 2007). However, in spite of the plausibility of this effect, Woodley (2012b) argued, that the strength and pace of the IQ gains observed over the years would be too powerful to be explained by hybrid vigor effects, since those would not be fast or strong enough to explain those gains.

Steen (2009) showed, that through environmental exposure to lead, blood lead levels increase, which has been linked to impairments of neural development, especially in young

children. When looking at IQ gains associated with the Flynn effect, it has been found, that the banning of lead paint and lead gasoline in the United States are correlated with increased intelligence test performance (Kaufman et al., 2014; Nevin, 2000). Nevertheless, the observation that IQ gains were stronger in adults than in children cannot be sufficiently explained by this theory as well as the notion that IQ gains have been slowing down following lead bans (Pietschnig & Voracek, 2015).

A theory, which has been proposed by Storfer in 1999 is based on a mechanism called genomic imprinting, which describes the influence of environmental conditions on reproduction information and genetic expression. Thus, it would be plausible, that visually stimulating environments would lead to changes in the reproduction information, therefore manifesting as increased cognitive abilities of the offspring. Because of difficulties testing that hypothesis as well as fluctuations of the IQ gains, it has been found that this theory would not be able to explain the Flynn effect fully (Pietschnig & Voracek, 2015).

1.4 The current study

In most of the proposed explanations for the Flynn effect it has been assumed that the IQ changes over the years have emerged due to a differentiation of knowledge, yet domain differences cannot be sufficiently explained with this approach. Thus, instead of focusing on a specification of knowledge, the goal of this study is to examine the intelligence test measures themselves in regards to their items and their contents. The idea for this hypothesis was already postulated by Rodgers (1998), who argued that intelligence tests would most likely get changed over the years, thus perhaps leading to IQ changes over time. An increase in intelligence test performance may be plausibly explained by an acculturation of knowledge from one generation to the other, facilitating answering items that have been affected by this process. Hence, people would find it easier to answer items from older test versions that have become part of the cultural knowledge but have not been at the time when the test was developed.

Similar effects of cultural changes over generations would emerge, if items include knowledge, that becomes more common over the years, thus again making it easier for people to answer those items (Flynn, 2007).

Interestingly, also an opposing effect could appear through these changes in intelligence tests over the years, leading to a negative Flynn effect. This would happen, if items become obsolete after some time, making it difficult for younger generations to answer those items. Thus, a person who would complete two versions of the same intelligence test

battery that have been normed in different years, would probably achieve a higher IQ score on the newer test version versus the older one, if item-obsolescence was present. Methodically, every item of each test version would have to be examined and all of the changed items from one test version to the other would have to be administered. The manifestation of item-obsolescence would then lead to a better performance on items from the new test version, as for example when an item from the older version cannot be properly understood in younger years. An improvement of factual knowledge oftentimes leads to these outdated items, thus items that have been correct in older generations now have lost their meaning or have become factually incorrect. Knowledge-based questions might also include facts that lose their relevance or validity over the years, for example when asking about the current president or population density of a specific city. Younger generations might know the current president or population density of that city, but not that of decades before that, where the items have been created and standardized, hence achieving lower IQ points on subscales with obsolete items, as such.

The changes in difficulties of items over time differ between fluid and crystallized intelligence, leading to biased IQ comparisons. For instance, the effects of obsolete items would mostly affect crystallized subtests and therefore also the crystallized intelligence test performance in the population. That is, because items of subtests testing crystallized intelligence get changed more often during test revisions due to language and knowledge changes in the population, rather than items of subtests that test fluid intelligence, which mostly stay the same or similar due to the fact that no prior knowledge is required to solve these items. The worse performance of individuals in tests with norms that have been established in earlier years would then result in a negative or at least lower Flynn effect in crystallized intelligence, showing a seemingly decreasing intelligence test performance over generations. However, considering the effect of item obsolescence, leading to an increased difficulty of items in older test versions, the real changes of crystallized intelligence are being masked. Consequently, the domain specific Flynn effect might be misjudged for crystallized intelligence due to the outdated items in older test norms, making the gap between fluid and crystallized intelligence seem bigger than it actually would be.

Consequently, there has to be another reason for the Flynn effect in the domain of fluid intelligence. Considering intelligence test norm changes over the years, not only the norm sample consists of a diverse group of people, who might differ in intelligence in general, but also the transformation of raw scores to standardized IQ scores differs in each norm revision of the Hamburg Wechsler Intelligence Test. Hence, when transforming the raw scores into IQ

scores it might seem as though younger generations generally have higher IQ's than older generations, due to the assignment of lower IQ scores in newer test versions to same raw scores as in older tests. Because older generations' raw scores were calculated using older norm tables, their IQ scores then seem lower. In short, the same raw score in an older and newer test version would lead to different IQ values – lower values in newer test versions and higher values in older test versions. This would happen due to higher IQ scores of the norm sample. The resulting positive Flynn effect in the fluid intelligence domain therefore reflects an increased performance of the norm sample, which was used to create these norm tables in the first place.

Concluding, the negative Flynn effect in the crystallized intelligence domain might result of a better intelligence test performance in revised test versions in crystallized subscales, since those do not include obsolete items. Consequently, older test versions, that do include obsolete items, lead to worse performance in younger generations, who's actual knowledge and intelligence are being masked by the item obsolescence of unrevised test versions. This in turn leads to the observation, that those younger generations are seemingly less intelligent than the norm sample of the original test versions, that include the obsolete items.

To investigate, if this might be the case in the domain-specific Flynn effect, two versions of a well-established German intelligence test battery, specifically three subscales of the original Hamburg-Wechsler-Intelligence-Inventory (HAWIE) and its revision (HAWIE-R) were used in the current study. The original test inventory was published in 1956 (HAWIE; Hardesty & Lauber, 1956), whereas its revision came out 35 years after, namely in 1991 (HAWIE-R; Tewes, 1991), revealing changes and additions to the item pools, although most of the items stayed the same (Satzger et al., 1996). As Satzger et al. (1996) summarizes, a shift of test norms for the revised Wechsler Adult Intelligence Scale (WAIS-R; Wechsler, 1981) compared to older versions of the Wechsler Scales, which make use of older test norms, was found, overestimating test-takers intelligence when using those obsolete test norms. Following that logic, when comparing the original Hamburg Wechsler Intelligence Test from 1956 (HAWIE) to its revision from 1991 (HAWIE-R), it is expected, that lower IQ's will be found in the revised version of the HAWIE, as this test will most likely not overestimate intelligence levels of individuals due to outdated items compared to the older test version with the assumed obsolete test norms.

Thus, as mentioned above, in the current study, three subscales of the original as well as the revised Hamburg Wechsler Intelligence Test Battery (HAWIE & HAWIE-R) were used to examine if improved intelligence test performance in the context of a domain specific

Flynn effect might be due to item obsolescence, testing fluid ("Arithmetics"), crystallized ("Vocabulary") and fluid-crystallized ("Comprehension") intelligence, in order to see if there will be a difference in IQ points across subscales and test versions. Subsequently, the aim of this study is, first, to examine if there is a Flynn effect, manifesting in higher IQ scores on the original test (HAWIE) than on its revision (HAWIE-R), and secondly, to investigate if there is a domain-specific Flynn effect for fluid intelligence, which can be confirmed if the effect is higher in the subtest testing Arithmetic's ("Rechnerisches Denken") than in the Vocabulary ("Wortschatztest") or Comprehension ("Allgemeines Verständnis") subtests, since it primarily tests fluid-intelligence, rather than crystallized intelligence or both.

Consequently, the first hypothesis in the current study states, that there will be a significant IQ difference between the original test (HAWIE) and its revision (HAWIE-R), indicating the presence of a Flynn effect. Better performance on the original test would demonstrate a positive Flynn effect, whereas better performance on the revision would signal a negative Flynn effect, meaning that intelligence test performance declined in the general population over the period of 35 years.

Based on the assumption that some items and subscales may be affected by obsolescence, the second hypothesis of this study focusses on the influence of that item obsolescence on each subscale, thus assuming that there will be a significant raw score difference between items that have been changed from its original version in the HAWIE to the revised test (HAWIE-R). This would indicate that certain items might have become outdated over the years, thus causing different performance from one test version to the other in these specific items. For that matter it is necessary to carry out a categorization of items into the clusters "equivalent", "assumed equivalent" and "different" depending on whether items have been completely exchanged, slightly altered or stayed the same from one test version to the other. A differentiation of the Flynn effect per subscale depending on changed obsolete items is then possible to assess. Since it would be harder for individuals to solve items from the original test version if those items are affected by obsolescence, weaker effects of the Flynn effect or even a negative Flynn effect would be expected in subscales with a higher number of changed items.

The hypotheses as well as the methods used in this study were registered in the "Open Science Framework" (OSF; https://osf.io/jvqdr) prior to data collection to guarantee transparency.

2. Methods

2.1 Participants

The sample for this study was acquired through personal contact and consisted of mainly friends, family members or colleagues. All participants gave their written consent to participation and were accordingly informed about the procedure of the study. The option to prematurely quit the participation without a given reason was presented to all participants. After the completion of the study procedure, participants were debriefed about the goal of the study. There was a total sample size of N = 100 participants (54f; M = 26.3 years, SD = 7.9), all of which were German speaking individuals. 7% stated compulsory education as their highest level of education, 52% indicated the A-Levels as highest level of education, 27% held a Bachelor's degree and 14% had achieved a Master's degree or higher. None of the participants were excluded from the study.

2.2 Materials

Three different subscales of the Hamburg Wechsler Intelligence Test for Adults (HAWIE; Hardesty & Lauber, 1956) as well as its revised version, the Hamburg Wechsler Intelligence Test for Adults Revision (HAWIE-R; Tewes, 1991), were administered to all participants in this study. The subscales, namely the Arithmetic subscale ("Rechnerisches Denken"), the Vocabulary subscale ("Wortschatz-Test") and the Comprehension subscale ("Allgemeines Verständnis"), were chosen due to their relevance in the domain specificity of the Flynn effect, hence each chosen subscale measured a domain of intelligence or both. In this case, the Arithmetic subscale is an indicator for fluid intelligence, the Vocabulary subscale indicates crystallized intelligence and lastly, the Comprehension subscale measured both, fluid and crystallized intelligence, since it requires the use of prior knowledge and fluid components at the same time. The inner consistency of the test was determined by Cronbach Alpha and lies between α = .73 and .94, depending on the subscale.

2.2.1 Arithmetic

The Arithmetic subtest measures quantitative reasoning, concentration and mental manipulation, thus relying on working memory capacities. As illustrated above, this subtest is an indicator for fluid intelligence, which describes the ability to solve reasoning problems and therefore does not rely on the use of prior knowledge. In the Arithmetic subtest, participants are given orally administered arithmetic word problems, e.g. "An apple costs 25 cents. How much do six apples cost?", which they have to solve without the use of a calculator or any other aid. In the current study, participants were presented with 19 items, five of which were

identical or assumed equivalent in both tests, meaning that only small changes were made in the revision that did not alter the core meaning of the item. To avoid double administration of identical, non-replaced items, they were only administered once, but used twice to obtain IQ scores in the subsequent statistical analysis, as they were present in both tests, the HAWIE and the HAWIE-R. Five items were taken from the HAWIE, that did not appear in the HAWIE-R, and nine items were taken from the HAWIE-R, that did not appear in the HAWIE. There was a time limit for each item, depending on the difficulty, participants had to solve the items within a set time frame of 15, 30, 60, or 120 seconds. For each correctly solved item within that time frame a point was awarded, making up a total of 19 points if every item was solved.

2.2.2 Vocabulary

The Vocabulary subtest is a test that focuses mainly on measuring verbal comprehension, but also verbal expression and semantic knowledge are assessed. Hence, it is a subtest that tests for crystallized intelligence, the domain of intelligence holding knowledge which is acquired throughout life. Therefore, solving the items presented in the Vocabulary subtest requires the use of prior knowledge. Participants are orally instructed to define the meaning of words that are presented to them, e.g. “What does the word *apple* mean?”. In the current study, 50 words were presented, 24 of which were equivalent in both versions of the HAWIE, the original test as well as its revision, 18 words were taken from the HAWIE, and 8 were taken from the revised version (HAWIE-R), that did not appear in the other test. No time limit was set in this subtest; thus, participants were allowed to take as much time as they needed to complete each item. One point was given for each correctly defined word, making up a total of 50 points, if the correct meaning of every word was given.

2.2.3 Comprehension

The Comprehension subtest measures the ability to express abstract social conventions, rules and expressions and relies on verbal comprehension and expression. Participants get asked several questions about social situations or common concepts, e.g. “What would you do, if you would find a sealed envelope on the street, with a stamp and an address written on it?”. Thus, it can be categorized into both, fluid and crystallized intelligence, since it does require a basic understanding of social norms and situations, yet it also involves reasoning and problem-solving skills. In this study, 18 questions were provided orally, that had to be answered correctly. Five of these questions were equivalent or assumed equivalent in both tests, five were administered from the HAWIE – those that were not present in the HAWIE-R, and eight items were administered from the HAWIE-R, that were not present in the HAWIE.

That way, all items, from both tests, the HAWIE and its revision, were presented to each participant. In this subtest, participants were able to achieve two points per item, depending on the accuracy of their answers. For the question "Why do clothes have to be washed?" for instance, one point was given when the answer indicated, that clothes have to be washed in order to "look good" or other subordinate reasons. Two points were granted, if the answer included a main reason for the question, such as hygiene in this particular example. In total, it was possible to reach 36 points, if all items were completed correctly and subsequently two points were achieved per question.

2.3 Procedure

Testing took place in the private rooms of the test administrator as well as in rooms of the University of Vienna, where the test was individually administered. In order to avoid influences of subscale order or standardization year, six different test manuals were created, to which participants were randomly assigned to. Subsequently, approximately half of the participants started each subtest with items from the HAWIE, followed by the items from its revision, the HAWIE-R, whereas the other half completed the test the other way around – starting the subtests with items from the HAWIE-R, followed by items of the original test (HAWIE). One third out of all the participants started the test with the Comprehension subtest ("Allgemeines Verständnis"), followed by the Vocabulary subtest ("Wortschatz-Test"), and finishing off with the Arithmetic subtest ("Rechnerisches Denken"). Another third started with the Vocabulary subtest, followed by Arithmetic, and finishing off with the Comprehension subtest. Finally, the last third started the test with the Arithmetic subtest, followed by the Comprehension subtest, and finishing off with Vocabulary.

Beforehand, all items were assigned to one of three categories, "equivalent", "assumed equivalent", or "different", depending on whether they were replaced in the revision, slightly changed or stayed the same. This was relevant for the following statistical calculations to differentiate between items from the older and the newer test version.

In order to investigate, whether there is item obsolescence present in the Flynn effect, all items that were changed from the original version (HAWIE) to the revised version (HAWIE-R) had to be identified in order to compare the test performance of participants on these items. In the Comprehension subscale there was a total of eight different items in the revised test version to the original one. In the Vocabulary subscale also eight items were changed, and in the Arithmetic subscale nine items were modified. For the analysis raw scores were

transformed into relative frequencies, since comparison would not have been possible otherwise, due to the different number of changed items.

Instructions, as well as all of the items, were given verbally by the test administrator. All of the answers given by the participants were written down into a test protocol for each participant, subtest and item, including the points achieved for every item. There was no time limit for the participants, but testing took approximately 30 to 45 minutes. After completion of the test, each participant was thanked and debriefed.

3. Results

For analysis, participants answers were coded into raw scores for each item and subtest according to the respective manual of each test version and subsequently converted into IQ scores by using the corresponding norm table for each subscale and test version. To perform all statistical calculations, the statistical software SPSS Statistics was used. Table 1 provides descriptive statistics for all subscales and test versions.

Table 1

Mean IQ and standard deviations of participants in subtests.

	N	Minimum	Maximum	Mean	SD
Comprehension (R)	100	80.00	145.00	117.20	13.20
Vocabulary (R)	100	85.00	145.00	118.55	10.90
Arithmetic (R)	100	85.00	105.00	96.40	6.66
Comprehension (H)	100	80.00	125.00	115.95	8.12
Vocabulary (H)	100	90.00	125.00	110.70	7.85
Arithmetic (H)	100	70.00	105.00	98.05	7.31

Note. R = revised test version (HAWIE-R); H =original test version (HAWIE).

In order to investigate the first hypothesis, that there will be a significant IQ difference between the original test (HAWIE) and its revision (HAWIE-R), a two-factorial repeated-measures analysis of variance (ANOVA) with 2x3 factors on IQ scores was calculated. The factors included the test version (subscales from the original vs. the revised subscales) as well as the three subscales (Arithmetic vs. Vocabulary vs. Comprehension subscales), which were all entered as within-subject factors. Due to lacking sphericity across the data, the analysis of variance was calculated using the Greenhouse-Geisser correction to adjust the lack of sphericity. There were five outliers across the different subscales and test versions, one in the

Comprehension subtest of the revised test (HAWIE-R) as well as one in the original test (HAWIE), one in the Vocabulary subtest of the revised test, and two in the Arithmetic subscale of the original test, as illustrated in Figure 1.

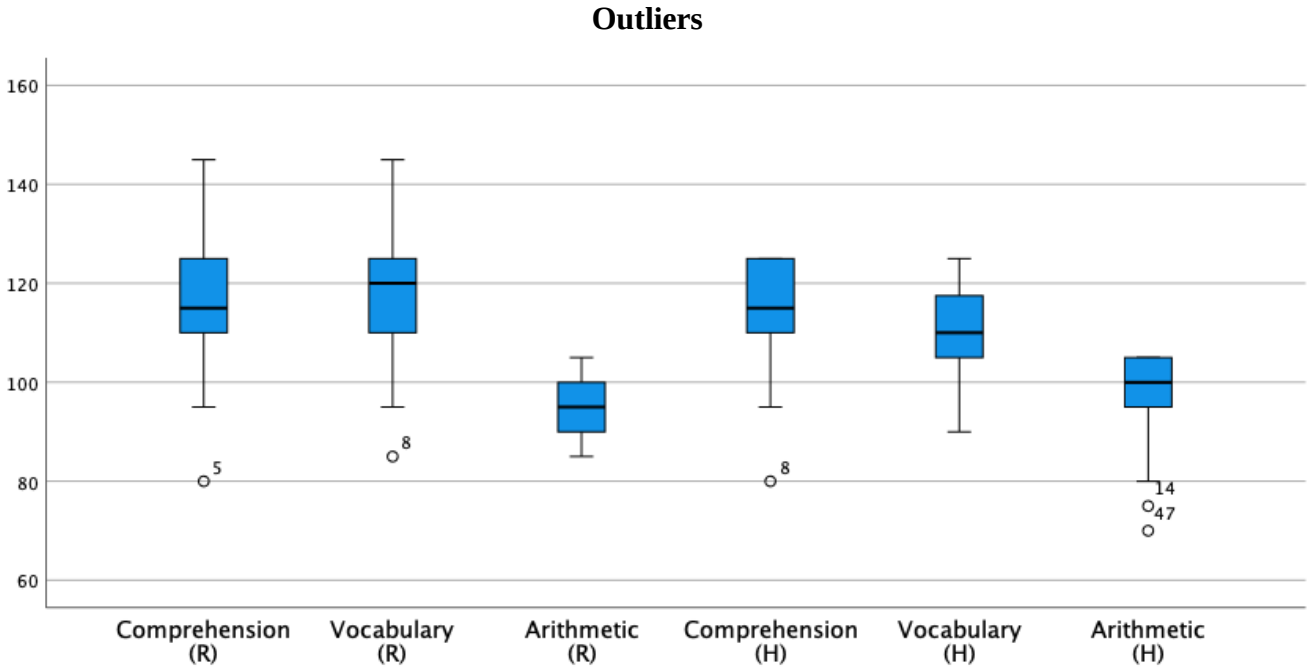

Fig. 1 Outliers in each subscale

Note. R = revised test version (HAWIE-R); H =original test version (HAWIE).

The results of the two-factorial repeated-measures ANOVA revealed significant main effects for test version and subtest ($p < .05$), as well as interaction effects for test version with subscale ($p < .05$), indicating that the test performance of participants differed on the original test versus the revised version across subscales (Table 2).

To further examine the direction of this effect, a simple effects analysis based on the estimated marginal means was carried out. This analysis is used in order to reveal the degree to which one factor is differentially effective at each combination of levels of the other factors. Results showed significant mean differences between the test versions with higher scores on the revised Hamburg-Wechsler Intelligence Test (HAWIE-R), with a mean difference of 2.48 IQ points ($p < .05$). Across the two test versions, the analysis revealed a significant mean difference of 7.85 IQ points in the Vocabulary subscale as well as a significant mean difference of -1.65 IQ points in the Arithmetic subscale ($p < .05$). There was no significant mean difference present in the Comprehension subtest.

In this case, a positive mean difference indicates better performance in the revised test version (HAWIE-R), thus suggesting a negative Flynn effect, whereas a negative mean

difference indicates better performance in the original test version (HAWIE), hence pointing towards a positive Flynn effect.

Table 2

Analysis of variance.

	F	*df*	*p*
Standardization year	35.73	1.00	<.001
Subscale	251.30	1.88	<.001
Standardization year * Subscale	42.82	1.69	<.001

Furthermore, to test the second hypothesis, that there will be a significant score difference between changed items, raw scores were transformed into relative frequencies according to the test version in order to compare the means of those items that were changed from the original test version to its revision. This was necessary because the number of changed items differed from one test version to the other. As illustrated in Table 3, the mean of the relative frequency of changed items in the Comprehension subscale was slightly larger in the original test (HAWIE) compared to its revision (.88 vs .86), similar to the means of the relative frequency of changed items in the Arithmetic subscale, where also a larger mean can be observed in the original test version (.88 vs .87). The exact opposite was the case when comparing the relative frequencies of changed items for the Vocabulary subscale, where the mean of changed items was significantly larger in the revised test version of the HAWIE (HAWIE-R; .88 vs .76).

Table 3

Descriptive statistics of relative frequencies for changed items.

	Mean	*N*	*SD*	*Standard error of mean*
Comprehension (H)	.88	100	.13	.01
Comprehension (R)	.86	100	.14	.01
Vocabulary (H)	.75	100	.10	.01
Vocabulary (R)	.87	100	.17	.01
Arithemtic (H)	.87	100	.18	.01
Arithmetic (R)	.86	100	.12	.01

Note. R = revised test version (HAWIE-R); H =original test version (HAWIE).

Subsequently, t-tests for paired samples were calculated, comparing the means of changed items from the original test to its revision across all three subscales. Results showed, that significance was only reached for the Vocabulary subscale ($p < .05$), indicating a significant change of mean performance for the respective subscale (Table 4).

Table 4

T-test with paired samples for changed items.

	T	*df*	*p*	*Cohen's d*	*Hedges Correction*
Comprehension H-R	1.38	99	.17	.13	.13
Vocabulary H-R	-7.57	99	<.001	-.75	-.75
Arithemtic H-R	.59	99	.55	.06	.05

Note. R = revised test version (HAWIE-R); H =original test version (HAWIE).

To further clarify these findings, Cohens effect sizes were calculated for the relative frequencies of changed items for each subscale, where an effect size of $d = 0.2$ indicates a small effect, an effect size of $d = 0.5$ is considered a medium effect, and an effect size of $d = 0.8$ shows a large effect (Cohen, 1988). The results of the calculation suggest a medium to large effect for the changed items in the Vocabulary subscale ($d = 0.75$) and no effect for the other two subscales ($d < 0.2$), as shown in Table 4. This reveals that in the Vocabulary subtest performance was better in the items from the revised test version, with an effect size of $d = 0.75$, which corresponds to a medium to large effect. This means, that participants scored higher on the items that were changed from the original test (HAWIE) to the revision (HAWIE-R), thus achieving a higher score in the revised test version (HAWIE-R). Nevertheless, when looking at the mean IQ difference between performance on the original and the revised test version, a decline of 7.85 IQ points can be observed in the Vocabulary subscale.

Furthermore, the mean IQ scores across the three subscales Arithmetic, Vocabulary and Comprehension differed from one test version to the other, yet only revealing significantly higher scores for the revised test in the Vocabulary subscale. The mean IQ scores in the Comprehension test were higher for the revised test version, whereas in the Arithmetic subscale the effect was the opposite, showing higher mean IQ scores for the original test version (Figure 2). Nevertheless, due to the small effect, both effects were negligible.

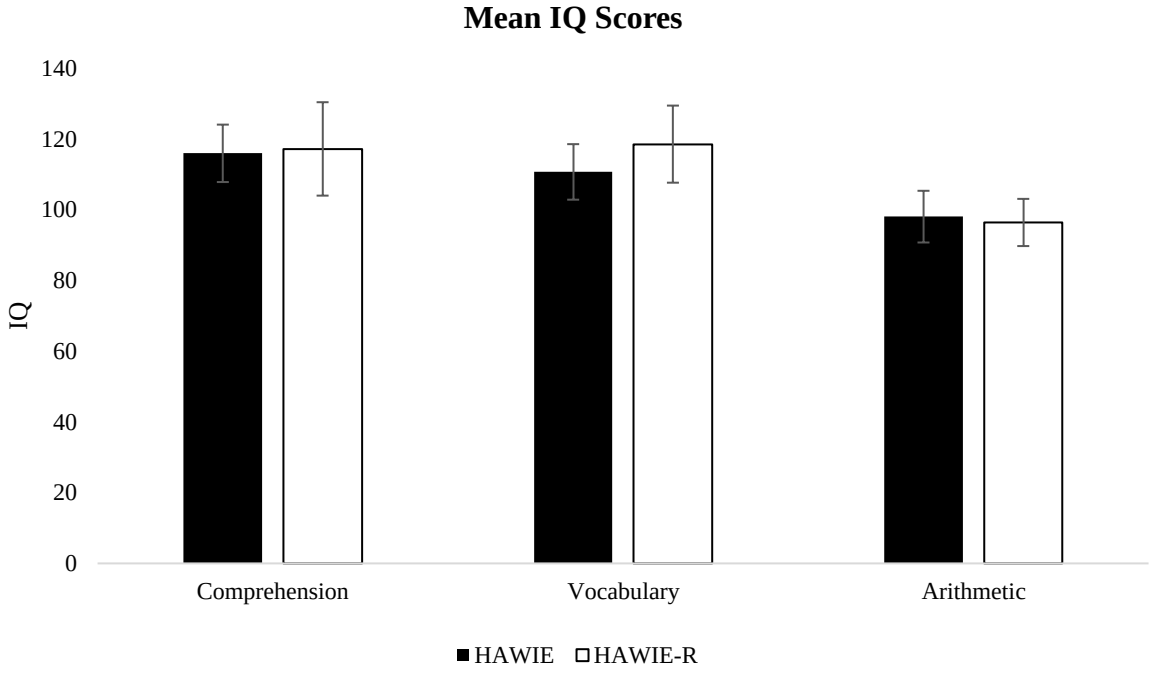

Fig. 2 Mean IQ on the original and revised subscales.

Finally, as a measure of the Flynn effect, the Delta IQ was computed, which describes the change of IQ points over a period of ten years. Similarly, results indicated the strongest effect for the Vocabulary subscale, showing gains of -2.24 IQ points per decade for the revised test version and little to no effects for the Comprehension subscale with a Delta IQ of +0.35. Additionally, (nonsignificant) gains of +0.47 IQ points per decade for the original test were observed in the Arithmetic subscale, indicating a positive Flynn effect.

4. Discussion

The goal of this study was to investigate the reasons behind a domain-specific Flynn effect, specifically stronger IQ gains in fluid intelligence compared to crystallized intelligence. Although previous studies in the field of intelligence and the Flynn effect set their focus on finding plausible explanations for the Flynn effect in general, domain specificity has been largely neglected.

Consequently, the influence of better education for example may play a role in the observed increase of intelligence test performance, however, it does not account for stronger gains in fluid intelligence. Following that reasoning, the focus of the current study was to investigate the influence of item obsolescence due to test norm changes on intelligence test performance in fluid and crystallized intelligence. This was done with a sample of 100

participants, who completed three subscales of two different versions of the Hamburg Wechsler Intelligence Test for Adults (HAWIE & HAWIE-R).

In order to test the equivalence of the test norms of both, the HAWIE and the HAWIE-R, Satzger et al. (1996) conducted a study, where participants were presented with all subscales of the original test (HAWIE), as well as those from the revision (HAWIE-R), leaving out identical items to prevent fatigue effects. Results of that study showed, that the points from each subscale of the old versus the new test correlated highly, indicating that both test versions actually tested the same skills (Satzger et al., 1996). Differences among the two versions of the Hamburg Wechsler Scales were the distribution of the general IQ, which was broader in the revised version of the Wechsler test (HAWIE-R), secondly, the mean IQ, which was higher in the original test (HAWIE) compared to its revision, and lastly, the IQ points in each subscale, showing higher IQ scores in the subscales "Allgemeines Verständnis" ("Comprehension"), "Rechnerisches Denken" ("Arithmetic"), and "Wortschatztest" ("Vocabulary") in the revised test version (HAWIE-R), suggesting a decrease in the performance level of these skills in the general population. Similar findings were reported by Schallberger (1987), who also found decreasing performance in arithmetic tasks.

Satzger et al. (1996) as well as Priester and Kukulka (1958), Priester and Kerekjarto (1960) and Tewes (1991) found a high inner consistency for both test versions, indicating equivalent contents in both tests. Therefore, in the context of norm changes in IQ tests, a lower general IQ in the revised Wechsler Scale (HAWIE-R) implies, that the intelligence level of the general population increased. These findings are especially important for the current research, since they serve as a basic requirement in order to compare the two test versions in the first place.

Regarding the expected results, the findings in this study confirmed the presence of a Flynn effect, which was observed in both directions – negative as well as positive, due to influences of item-obsolescence. Specifically, the data indicated a significant IQ difference between the performance on the original test (HAWIE) versus the revised test version (HAWIE-R). The direction of this effect was further examined by computing an analysis of simple effects of each subscale as well as the interaction between subscale and test version, where a positive Flynn effect in the Arithmetic subscale was noticeable, hence pointing towards a domain specific Flynn effect in fluid intelligence. In the current data, this is visible through a better mean performance in the original test version (HAWIE) compared to its revision. This was also found in past research, as for example in a meta-analysis conducted by Pietschnig and Voracek (2015), who discovered a domain specific Flynn effect in fluid

intelligence of 0.22 IQ points per year. Nevertheless, in the current study the mean performance in the Arithmetic subtest, which measures fluid intelligence showed only slight IQ differences of 1.65 IQ pints between the years 1956 and 1991, which equals a gain of 0.05 IQ points each year. This finding could be explained by the observed decline of the Flynn effect in recent decades, as shown in several studies (e.g., Dutton & Lynn, 2015). Since for the current analysis a sample from the year 2020 was used, this smaller gain could be interpreted as further changes in the Flynn effect over the years and might be an indicator for a possible reversal of the Flynn effect, as past literature already suggested (Pietschnig & Voracek, 2015). However, this is only an assumption and was not specifically looked at in this study.

There were no IQ gains in the Vocabulary subscale, as predicted by the hypothesis that stronger gains are found in fluid, but not crystallized intelligence. In fact, a decrease of intelligence test performance of 7.85 IQ points over a time span of 35 years was observed, which corresponds to a decline of 0.22 IQ points per year. Consequently, in the Vocabulary subscale participants scored higher on the revised test version (HAWIE-R) compared to the original one, resulting in a negative Flynn effect. Lastly, in the Comprehension subscale, which includes fluid as well as crystallized elements, no significant mean differences in test performance could be observed, which fits the expectations of a domain specific Flynn effect.

These findings generally support prior research regarding the Flynn effect, nevertheless, the decrease of intelligence test performance in the Vocabulary subscale, thus in the domain of crystallized intelligence, has not been reported so far. A recent thesis, done by Höltl (2019) at the University of Vienna, who also examined the role of item obsolescence in the domain specific Flynn effect, came to a similar conclusion. In that study, a decline of 0.36 IQ points per year over a period of 30 years in the domain of crystallized intelligence could be found, hence supporting the findings presented in the current study.

In order to investigate, whether there is item obsolescence present in the Flynn effect, all changed items from the HAWIE to the HAWIE-R were identified and later analyzed, revealing better performance in items from the revised test version in the Vocabulary subtest. However, the mean IQ difference between performance on the original and revised test version shows a decline of 7.85 IQ points in the Vocabulary subscale. Ultimately, this shows, that items that became obsolete do have an effect on the overall intelligence test performance and might be masking the actual intelligence test performance in crystallized intelligence. The increase in performance over the years might therefore be underestimated, hence creating a seemingly bigger gap between fluid and crystallized intelligence test performance than

actually present, since participants have a hard time solving obsolete items and therefore achieve less IQ points in tasks involving these obsolete items. Although the actual performance is better, the item obsolescence masks this through a lower IQ score in general.

This reasoning is appropriate for all scales that make use of crystallized items, since these are the ones typically affected by item obsolescence compared to fluid items. To illustrate this, one has to understand that crystallized tests comprise items testing one's knowledge, thus relying on prior memories and acquired facts (Pietschnig & Voracek, 2015). Thus, questions that are mainly knowledge-based might include facts that lose their relevance or validity over the years. This would lead to individuals achieving lower IQ points on subscales with obsolete items.

In contrast, fluid intelligence does not require the use of prior knowledge that might become obsolete, since fluid items are solvable using only logic and reasoning. In the fluid subscale used in this study for example, participants had to solve arithmetical problems. Since the strategy that has to be used in order to solve these problems does not vary depending on the standardization year of the items, no item obsolescence occurs.

As expected, possible item obsolescence could only be found in the Vocabulary subscale, which lines up with the hypothesis, that only crystallized items are affected by outdated information in intelligence testing. Hence, the analysis of the items that were changed from the original test version (HAWIE) to its revision (HAWIE-R) showed nonsignificant effect sizes of $d = .14$ for the Comprehension subscale, and $d = .06$ for the Arithmetic subscale. Nevertheless, according to the theory of item obsolescence in crystallized items of intelligence tests, it can be noticed, that in the Comprehension subscale, which includes both, fluid, as well as crystallized components, effect sizes were slightly higher than in the Arithmetic subscale, which mostly assesses fluid intelligence. This tendency towards a small effect ($d = 0.2$) in the Comprehension subscale could be explained by the crystallized portions of the items included in this subtest, which might be affected by item obsolescence. Consequently, the Arithmetic subscale, which is most likely not affected by item obsolescence, due to its property to mostly measure fluid intelligence, shows no effect at all when comparing only the modified items from the older test version to the more recent one.

The findings in this study can at least partly confirm the hypotheses stated at the beginning of this thesis, since there were significant IQ differences observable when comparing intelligence test performance of participants in the Hamburg Wechsler Intelligence Test for Adults (HAWIE) and its revised version (HAWIE-R), which was established 35 years after the original test version. Furthermore, score differences in items that were changed

from the original test to the revision were noticeable. However, these score differences between changed items showed significance only for the Vocabulary subscale, therefore only confirming parts of the second hypothesis made in this study.

As for the Flynn effect in general, the results were in line with current research, especially regarding a stagnation and possible reversal of the effect. For instance, a decline of intelligence was already noticed in France between the years 1999 and 2009, where particularly in the domain of crystallized intelligence a stronger decrease of IQ points was observed than in other intelligence domains, with a reduction of 4 IQ points over the span of 10 years (Dutton & Lynn, 2015). Similarly, findings also showed reduced IQ scores in Finland, with stronger declines in crystallized subtests as well (Dutton & Lynn, 2013).

Based on the findings in the current study, the domain differences in the Flynn effect discovered in past research might also have been influenced by item obsolescence, thus making it seem as though the gap of declining IQ points achieved in different subtests in intelligence tests between crystallized and fluid intelligence is bigger than it actually is.

The reasons as to why the Flynn effect is potentially changing its direction are manyfold. Theories concerning this matter range from the influence of immigration as a potential factor for the reversal of the Flynn effect, to mortality and sex differences. Taking a closer look at these postulations, a study conducted by Lynn and Vanhanen (2002) argues, that due to increased immigration in western Europe the general intelligence test performance of the population decreases, possibly explaining a reversal of the Flynn effect. This reasoning stems from studies, that indicate a mean IQ difference of 10 to 15 IQ points between north African and southwest Asian immigrants compared to the European population (Lynn & Vanhanen, 2002). Similarly, a Danish study recorded an increase in immigration of 350.000 people from the years 1980 to 2012, with a mean IQ difference of almost 14 IQ points between immigrants and the Danish population (Kirkegaard, 2013). However, recent evidence dismantles the hypothesis of immigration as a possible explanation for a decline of general intelligence and thus a reversal of the Flynn effect, as for instance a study conducted by Pietschnig et al. (2018) shows. The researchers involved in this analysis were not able to find negative effects of immigration on the general IQ of the population, but rather found positive influences of immigration on intelligence test performance. Nevertheless, whether positive or negative effects of intelligence test performance of immigrants are observed, due to their vast distribution an influence on the general IQ levels of a specific country would be questionable, as Dutton and Lynn (2013) state.

A different theory regarding a reversal of the Flynn effect consists of the effects of dysgenic factors on intelligence, which was recently explored in a meta-analysis (Pietschnig et al., 2018). This theory suggests, that dysgenic processes diminish cognitive abilities, thus leading to a reversed Flynn effect. However, Pietschnig et al. (2018) were not able to find consistent evidence supporting this theory, which should be able to show a negative correlation between fertility and mean IQ changes in the general population. In their meta-analysis though, these correlations could not be found, except for small effects in the domain of crystallized intelligence. Interestingly, in the domain of fluid intelligence, an opposite effect could be observed, showing a positive correlation between fertility and IQ changes (Pietschnig et al., 2018). Due to these inconsistencies in their study, further research in this regard is needed to draw a plausible conclusion for this theory.

Additionally, in their meta-analysis Pietschnig et al. (2018) examined the possible influence of reduced mortality on the reversal of the Flynn effect. According to this theory, the reproductive age is met by more and more individuals over the latest decades due to advances in the field of medicine, which leads to more individuals with a lower IQ to pass on their genes to the next generations and so on (Nyborg, 2012). Similar to the hypothesis of dysgenics as a possible factor for the observed reversal of the Flynn effect, also for the theory of mortality only inconsistent findings were observable (Pietschnig et al., 2018).

Ultimately, the distribution of men and women in the samples might influence generational IQ changes, since differences in performance have been noticed between sexes (Dutton, Linden, & Lynn, 2016). Yet, in a study conducted by Pietschnig et al. (2011) sex differences did not have a significant impact on the Flynn effect.

In the context of the current study inferences based on these potential causes for a stagnation and reversal of the Flynn effect cannot be made, as these influences have not been specifically controlled for. For instance, immigration as a potential factor for a reversal of the effect was not possible to assess in this study since data only included Austrian and German participants. Dysgenics as well as mortality were also not empirically tested in the current research. Further investigation in this field is needed in order to answer this question more clearly.

On the other hand, the potential factors leading to an incline in generational IQ changes might indeed also have an impact on the stagnation and perhaps the reversal of those changes in intelligence test performance. As mentioned earlier in this thesis, influences like better education or nutrition might have been responsible for at least some parts of the observed increases in IQ points over the decades, as Lynn (2009) and Ceci and Williams (1997)

postulate. However, these factors might also be partly responsible for a stagnation of the Flynn effect, due to ceiling effects. Following this reasoning, especially in developed countries the influence of factors like improved nutrition and education might have reached a level where no further improvement is possible over the last few years, thus slowing down the observed Flynn effect. Hence, due to the different speed of development across countries these effects might reach a ceiling at different times, leading to country-specific effects instead of a linear and stable effect over different world regions (Pietschnig & Gittler, 2015).

Moreover, the negative correlation between the Flynn effect and the g-factor of intelligence might explain a possible reversal of the Flynn effect even further. According to Pietschnig (2016), intelligence test performance only increases in specific intelligence domains, rather than the general intelligence itself, meaning that domains that are not highly correlated with the *g*-factor might increase, while the general intelligence factor stays the same. Thus, fluid intelligence, a domain that is not correlated to the *g*-factor, might increase, while other intelligence domains that are closely related to the *g*-factor might stagnate or decline. Combined with ceiling effects of factors that usually have led to an increase of intelligence test performance in certain intelligence domains, resulting in a stagnation or decline of IQ points in these domains, a reversal of the Flynn effect might become visible.

These influences might also play a role in the current findings, where a negative Flynn effect could be assessed in the domain of crystallized intelligence as well as the tendency towards a stagnation of intelligence test performance in fluid intelligence. A reversal of the IQ gains in crystallized subscales could be explained by item obsolescence in non-revised older tests, yet a stagnation of fluid or general intelligence cannot be explained by this effect. Thus, a combination of ceiling effects of factors that would usually account for a positive Flynn effect and a negative correlation between the Flynn effect and the g-factor of intelligence could be responsible for the stagnating effects in fluid intelligence observed in this study.

4.1 Limitations

In the current study the following limitations should be considered. First, only three subscales were administered to the participants in this study, one covering each intelligence domain. The use of other subscales, that also cover these domains or an additional administration of other subscales would be beneficial in future research in order to better generalize the results.

Moreover, it has been noted, that in the lower intelligence range the Flynn effect is more obvious than in higher IQ's (Lynn & Hampson,1986), yet in the current sample almost all participants scored an IQ of 100 or above in each subscale, thus leading to possible weak

Flynn effects in the subscales tested. A general IQ was not possible to calculate for each participant since only three subscales were administered and a minimum of six subscales is needed in order to determine a general IQ (HAWIE; Hardesty & Lauber, 1956).

5. Conclusion

Since the first observation of generational IQ changes in the 1900s (Flynn, 1984), many studies have been conducted concerning this topic of research to identify possible explanations for the global IQ gains over the years (Pietschnig & Voracek, 2015). However, a plausible reason for a domain specific Flynn effect has yet to be found, thus, domain differences of generational IQ gains in fluid and crystallized intelligence were the research focus in the current study. As the results reveal, a Flynn effect was found in fluid as well as crystallized intelligence, yet effects pointed in opposite directions. In mainly fluid subscales a positive Flynn effect was indicated, whereas in mostly crystallized subscales a negative Flynn effect was detectable. A potential reason for these seemingly big differences between fluid and crystallized intelligence regarding the Flynn effect might be the influence of item obsolescence. In older, unrevised intelligence test versions certain items might become outdated, since cultural knowledge develops over time, leading to younger generations being unable to solve these items correctly. Obsolete items can be usually found in crystallized subtests, because they reflect evolving knowledge rather than fluid subtests that only rely on abstract cognitive operations. Hence, the finding of a negative Flynn effect in crystallized subscales in this study could be explained by item obsolescence and thus masking actual IQ gains in crystallized intelligence. The domain specific differences are therefore exaggerated by item obsolescence, meaning that without this masking, maybe no or at least smaller differences would be present in the different intelligence domains.

Regarding fluid intelligence, a positive Flynn effect could be found. However, since the effect is small and pointing towards a stagnation of the Flynn effect this finding matches other research in the past years (e.g. Dutton, van der Linden, & Lynn, 2016) and supports the notion of a potential reversal of the generational IQ changes in the future. On this account, the present research functions as a great addition to past studies in this field (Pietschnig, Gittler, Höltl, Tran & Voracek, 2019) and might operate as a building block for future research concerning a reversed Flynn effect and domain differences.

Literature

Benson, E. S. (2003). Intelligent intelligence testing. *Monitor on Psychology*, *34*(2). http://www.apa.org/monitor/feb03/intelligent.html

Binet, A. & Simon, T. (1905). New methods for the diagnosis of the intellectual level of subnormals. In *The Development of Intelligence in Children (The Binet-Simon Scale)* (pp. 37–90). Leopold Classic Library.

Boake, C. (2002). From the Binet–Simon to the Wechsler–Bellevue: Tracing the History of Intelligence Testing. *Journal of Clinical and Experimental Neuropsychology*, *24*(3), 383–405. https://doi.org/10.1076/jcen.24.3.383.981

Brand, C. (1987). Keeping up with the times. *Nature*, *328*(6133), 761. https://doi.org/10.1038/328761a0

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Ceci, S. J. & Williams, W. M. (1997). Schooling, intelligence, and income. *American Psychologist*, *52*(10), 1051–1058. https://doi.org/10.1037/0003-066x.52.10.1051

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)* (2. Aufl.). Routledge.

Colom, R., Lluisfont, J. & Andrespueyo, 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

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. & 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

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

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. (2006). Tethering the elephant: Capital cases, IQ, and the Flynn effect. *Psychology, Public Policy, and Law*, *12*(2), 170–189. https://doi.org/10.1037/1076-8971.12.2.170

Flynn, J. R. (2007). *What is intelligence? Beyond the Flynn effect.* Cambridge University Press.

Hardesty, A., & Lauber, H. (1956). *Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE).* Huber.

Healy, W. (1914). A pictorial completion test. *Psychological Review*, *21*(3), 189–203. https://doi.org/10.1037/h0075712

Healy, W. (1921). Pictorial Completion Test II. *Journal of Applied Psychology*, *5*(3), 225–239. https://doi.org/10.1037/h0073697

Herrnstein, R. J. & Murray, C. (1996). *The Bell Curve: Intelligence and Class Structure in American Life (A Free Press Paperbacks Book)* (Illustrated Aufl.). Free Press.

Höltl, Franziska (2019). *Einflüsse von Itemobsoleszenz auf generationsspezifische Testnormverschiebungen.* Masterarbeit, Universität Wien. Fakultät für Psychologie. BetreuerIn: Pietschnig, Jakob

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. *Woodcock (Eds.), Woodcock-Johnson technical manual* (pp. 197–232). Riverside.

Hunt, E. (2011). *Human intelligence*. Cambridge University Press.

Jensen, A. R. (1998). *The g Factor: The Science of Mental Ability (Human Evolution, Behavior, and Intelligence)* (1St Edition). Praeger.

Kaufman, A. S., Zhou, X., Reynolds, M. R., Kaufman, N. L., Green, G. P. & Weiss, L. G. (2014). The possible societal impact of the decrease in U.S. blood lead levels on adult IQ. *Environmental Research*, *132*, 413–420. https://doi.org/10.1016/j.envres.2014.04.015

Kirkegaard, E. O. W. (2013). Predicting Immigrant IQ from their Countries of Origin, and Lynn's National IQs: A Case Study from Denmark. *Mankind Quarterly*, *54*(2), 151–167. https://doi.org/10.46469/mq.2013.54.2.2

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

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. & 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

Mcgrew, K. S. (1997). Analysis of the Major Intelligence Batteries According to a Proposed Comprehensive Gf-Gc Framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 151-179). The Guilford Press.

McGrew, K. (2005). The Cattell-Horn-Carroll Theory of Cognitive Abilities: Past, Present, and Future. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues*. The Guilford Press.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

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

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

Nevin, R. (2000). How Lead Exposure Relates to Temporal Changes in IQ, Violent Crime, and Unwed Pregnancy. *Environmental Research*, *83*(1), 1–22. https://doi.org/10.1006/enrs.1999.4045

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F. & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist, 67*(2), 130–159. https://doi.org/10.1037/a0026699

Nyborg, H. (2012). The decay of Western civilization: Double relaxed Darwinian Selection. *Personality and Individual Differences*, *53*(2), 118–125. https://doi.org/10.1016/j.paid.2011.02.031

Petermann, F. (2012). *WAIS-IV Wechsler Adult Intelligence Scale – Fourth Edition*. Pearson Assessment.

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

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

Pietschnig, J., Gittler, G., Höltl, F., Tran, U. S., & Voracek, M. (2019). Smaller Flynn Effects for crystallized intelligence may be rooted in item obsolescence: Results from archival data and a direct test of generational IQ gains. Twenthieth Annual Conference of the International Society for Intelligence Research (ISIR), 13. -15.07.2018, Minneapolis, MN.

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. (2011). Female Flynn effects: No sex differences in generational IQ gains. *Personality and Individual Differences*, *50*(5), 759–762. https://doi.org/10.1016/j.paid.2010.12.019

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*, *6*(2), 267-283.

Pintner, R., & Paterson, D.G. (1917). *A scale of performance tests.* Appleton.

Priester, H.J., Kukulka, R. (1958). Vergleichsuntersuchungen zum Hamburg-Wechsler-Intelligenztest für Kinder (HAWIK) und Binet-Bobertag und zum HAWIK und dem Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE) im Bezug auf die Intelligenzquotienten und die Benutzung dieser Tests als Paralleltests. *Diagnostica*, *4*, 6–16.

Priester, H.J., Kerekjarto, M. (1960). Weitere Forschungsergebnisse zum Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE) und zum Hamburg-Wechsler-Intelligenztest für Kinder (HAWIK). *Diagnostica*, *6*, 86–94.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* University of Chicago Press.

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

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

Rowe, D. C. & Rodgers, J. L. (2002). Expanding variance and the case of historical changes in IQ means: A critique of Dickens and Flynn (2001). *Psychological Review*, *109*(4), 759–763. https://doi.org/10.1037/0033-295x.109.4.759

Runquist, E. A. (1936). Intelligence test scores and school marks of high school seniors in 1929 and 1933. *School and Society, 43*, 301-304.

Satzger, W., Dragon, E., & Engel, R. R. (1996). Zur Normenäquivalenz von HAWIE-R und HAWIE [The equivalence of the original and revised German versions of the Wechsler Adult Intelligence Test]. *Diagnostica, 42*(2), 119–138.

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

Schallberger, U. (1987). HAWIK und HAWIK-R: Ein empirischer Vergleich [HAWIK and HAWIK-R: An empirical comparison]. *Diagnostica*, *33*(1), 1–13.

Spearman, C. (1927). The abilities of man. In *The abilities of man.* Macmillan.

Steen, G. R. (2009). *Human Intelligence and Medical Illness: Assessing the Flynn Effect (The Springer Series on Human Exceptionality)* (2009. Aufl.). Springer.

Sternberg, R. J. (1984). A contextualist view of the nature of intelligence. *International Journal of Psychology*, *19*(1–4), 307–334. https://doi.org/https://doi.org/10.1080/00207598408247535

Storfer, M. (1999). Myopia, Intelligence, and the Expanding Human Neocortex: Behavioral Influences and Evolutionary Implications. *International Journal of Neuroscience*, *98*(3–4), 153–276. https://doi.org/10.3109/00207459908997465

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

Terman, L. M. (1916). *The Measurement of Intelligence An Explanation of and a Complete Guide for the Use of the Stanford Revision and Extension of the Binet-Simon Intelligence Scale*. Houghton Mifflin.

Tewes, U. (1991). *HAWIE-R Hamburg-Wechsler-Intelligenztest für Erwachsene Revision 1991*. Huber.

Thorndike, R.M., & Lohman, D.F. (1990). *A century of ability testing*. Riverside.

Thurstone, L. L. (1935). The Vectors of Mind. In *The Vectors of Mind.* University of Chicago Press.

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

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

Wechsler, D. (1956). *Die Messung der Intelligenz Erwachsener*. Textband zum Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE); Deutsche Bearbeitung Anne von Hardesty, und Hans Lauber. [*The measurement of adult intelligence*.]. Hans Huber.

Wechsler, D. (1981). The psychometric tradition: Developing the Wechsler adult intelligence scale. *Contemporary Educational Psychology*, *6*(2), 82–85. https://doi.org/10.1016/0361-476x(81)90035-7

Wheeler, L. R. (1942). A comparative study of the intelligence of East Tennessee mountain children. *Journal of Educational Psychology*, *33*(5), 321–334. https://doi.org/10.1037/h0063294

Woodley, M. A. (2012). A life history model of the Lynn–Flynn effect. *Personality and Individual Differences*, *53*(2), 152–156. https://doi.org/10.1016/j.paid.2011.03.028

Yerkes, R.M. (Ed.) (1921). *Psychological examining in the United States Army. Memoirs of the National Academy of Sciences*, 15 (Parts 1-3). Government Printing Office.

Zajonc, R. B. & Mullally, P. R. (1997). Birth order: Reconciling conflicting effects. *American Psychologist*, *52*(7), 685–699. https://doi.org/10.1037/0003-066x.52.7.685

Appendix

Abstract

In the field of intelligence, the phenomenon of rising intelligence test performance over generations, that has become well known as the “Flynn effect”, is now widely researched. These increasing intelligence test performances are characterized by an incline of three to five IQ points per decade, with variations between countries and domains, which have been observable since the 1900s (Flynn, 1984). Interestingly, a stagnation and even reversal of this effect has been noticed in the last 30 years, especially affecting Scandinavian countries (Teasdale & Owen, 2005). What factors might cause the Flynn effect and its potential reversal has been examined by several researchers, nevertheless, a comprehensive explanation for a country- and domain specific Flynn effect has yet to be found. In the current study, the influence of test norm changes on different intelligence domains has been investigated, specifically focusing on the hypothesis that item obsolescence might be a potential factor contributing to the observed generational IQ changes. For this matter, a total of 100 participants (54f; M = 26.3 years, SD = 7.9) were administered three subscales of the Hamburg Wechsler Intelligence Test for Adults (HAWIE; Hardesty & Lauber, 1956) as well as its revision, the Hamburg Wechsler Intelligence Test for Adults Revised (HAWIE-R; Tewes, 1991). A two-way repeated-measures ANOVA showed significant differences between performances in the original and the revised test versions. Further analysis revealed significantly better performances of participants in the original test version for items in the “Arithmetic” subscale, thus suggesting a positive Flynn effect in the domain of fluid intelligence. In the “Vocabulary” subscale better performance in the revised test has been observed, indicating a negative Flynn effect in crystallized intelligence, meaning that ability seemingly decreased over time. These changes in performance were assessed by specifically looking at the item scores of the items that were changed from the original to the revised test version in order to determine if the replacement of items is of considerable importance regarding generational changes of intelligence test performance. The significantly higher performance in the revised test versions on crystallized items points towards the effects of item obsolescence on IQ changes, masking crystallized IQ gains and thus exaggerating domain specific IQ differences in the Flynn effect.

Zusammenfassung

Ein generationenübergreifender Anstieg durchschnittlicher IQ Werte in Intelligenztests ist ein Phänomen, das heute unter dem Namen „Flynn Effekt“ bekannt ist. Die intensive Forschung in diesem Bereich zeigt, dass der Effekt zwischen Ländern und Intelligenzdomänen variiert, der Anstieg jedoch drei bis fünf IQ Punkte pro Jahrzehnt beträgt (Flynn, 1984). Obwohl einige plausible Hypothesen für die Entstehung des Flynn Effektes untersucht wurden, konnte bislang noch keine schlüssige Erklärung für die stärkeren Zuwächse in fluider Intelligenz, im Gegensatz zu kristalliner Intelligenz, gefunden werden. Die vorliegende Studie beschäftigt sich demnach mit dem Einfluss der Testnormverschiebungen auf die beiden Intelligenzdomänen fluide und kristalline Intelligenz. Vor allem Itemobsoleszenz als eine plausible Erklärung für weniger starke IQ Zuwächse in kristalliner Intelligenz werden im Rahmen dieser Arbeit diskutiert. Hierfür wurden 100 Versuchspersonen (54f; M = 26.3 Jahre, SD = 7.9) rekrutiert, welchen drei Subskalen des Hamburg Wechsler Intelligenztest für Erwachsene (HAWIE; Hardesty & Lauber, 1956) und dessen Revision (HAWIE-R; Tewes, 1991) vorgelegt wurden. Eine zweifaktorielle ANOVA mit Messwiederholung konnte signifikante Unterschiede zwischen der Leistung in der originalen und der revidierten Version des Tests feststellen. Weitere Analysen zeigten bessere Leistung der Versuchspersonen in der originalen Version des Intelligenztests in der Subskala „Rechnerisches Denken“, was einem positiven Flynn Effekt in der fluiden Intelligenzdomäne entspricht. Ein umgekehrter Effekt konnte in der Subskala „Wortschatztest“ beobachtet werden. Hier zeigten Versuchspersonen eine bessere Leistung in der revidierten Version des HAWIE, was einem Abfall der Leistung über die Zeit entsprechen würde. Dieser Effekt, der in diesem Fall nur kristalline Items betrifft, würde für die Hypothese der Itemobsoleszenz sprechen und somit kristalline IQ Zuwächse über die Zeit maskieren und eine Überschätzung der domänenspezifischen Unterschiede im Flynn Effekt zur Folge haben.