Intelligence 86 (2021) 101544

Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell

The Flynn effect in Germanophone preschoolers (1996–2018): Small effects, erratic directions, and questionable interpretations

Jakob Pietschnig *, Pia Deimann, Nicole Hirschmann, Ursula Kastner-Koller

Department of Developmental and Educational Psychology, University of Vienna, Austria

ARTICLE INFO

Keywords:
Flynn effect
Developmental tests
Cross-temporal meta-analysis
Measurement invariance

ABSTRACT

Generational intelligence test score changes were predominantly positive over most of the past century. However, so far, only little is known about this so-called Flynn effect in children that have not yet been exposed to formal schooling. So far, the cross-temporal trajectory of performance changes on developmental tests is unclear. Here, we investigated test score changes in Germanophone preschoolers on six areas of the Viennese Developmental Test (VDT/WET). First, we used data of standardization samples ($N = 1630$) in Austria and Germany to calculate changes between 1996 and 2008.5. Subsequently, we used a cross-temporal meta-analytic approach to investigate another 22 independent samples ($k = 1251$) from 2001 to 2018. Examination of both raw score and latent mean changes yielded mostly non-significant and trivial changes in cognitive development between three standardizations. Only change scores of the most fluid developmental domain showed positive signs, thus conforming to prior observations of larger Flynn effect for fluid than for crystallized intelligence (maximum overall changes ranged from −1.44 to 0.78 IQ points per decade). Results of our cross-temporal analyses were largely consistent in signs with overall changes, but failed to reach nominal significance in all instances. Our findings indicate that there is no convincing evidence for a Flynn effect in cognitive development in three-to-six year-olds. These findings support the role of education as an important driver of test score gains. Future research needs to determine if such a pattern may be a precursor of a Flynn effect stagnation or even its reversal.

1. Introduction

The Flynn effect (i.e., commonly understood as generational IQ test score changes; Flynn, 1984, 1987) has been shown in a wide variety of different cognitive domains globally. These test score changes were demonstrated to have been predominantly positive over large parts of the 20th century, yielding gains of about 3 IQ points per decade from 1909 to 2013 (Pietschnig & Voracek, 2015). Typically, gains were larger for measures of fluid than of crystallized intelligence and they were found to be negatively related to psychometric *g* (Must, Must, & Raudik, 2003; Pietschnig & Voracek, 2015; Woodley & Meisenberg, 2013; but see Colom, Juan-Espinosa, & Garcia, 2001, for different findings).

However, the trajectory of these gains was shown to be non-linear, exhibiting strong increases over some periods that are interspersed with periods of considerably lower gains in all domains or even virtual stagnation in crystallized intelligence. Importantly, towards the end of the past century, the gains were evidenced to considerably diminish, thus leading to speculation of an impending stagnation or even a possible reversal of the Flynn effect. However, despite their decreasing strength, these gains were still ongoing for all domains (excepting crystallized IQ) on a global scale.

Notwithstanding, in the past years, evidence for negative test score changes has been increasingly accumulating, thus indicating that such a reversal may already have taken place in several countries all across Europe (e.g., Dutton, van der Linden, & Lynn, 2016). Some recent studies even suggest that we may be soon witnessing a similar reversal in North America as well (Dworak, 2019; Schroeder, 2019). However, these accounts represent interesting, yet uncorroborated, evidence that is in need of further examination. So far, meta-analytic evidence that is supported by longitudinal archival data, has shown that a positive Flynn effect had been ongoing in the United States and the United Kingdom until the early 2010s (O'Keefe & Rodgers, 2017; Trahan, Stuebing, Hiscock, & Fletcher, 2014).

Although the Flynn effect has been intensively investigated in adolescent and adult populations, so far, not as much is known about potential test score changes in developmental tests. To date, comparatively more evidence is available about test score changes in cognitive development. Some evidence suggests that, similar to IQ tests,

* Corresponding author at: Department of Developmental and Educational Psychology, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria.
E-mail address: jakob.pietschnig@univie.ac.at (J. Pietschnig).

https://doi.org/10.1016/j.intell.2021.101544
Received 27 July 2020; Received in revised form 22 April 2021; Accepted 23 April 2021
Available online 18 May 2021

developmental test scores may have been increasing over the past century. Results of UK-based samples of three-to-thirteen year-olds on the Draw-a-person test have steadily increased from 1902 to 1968 (Genovese, 2018). Other investigations estimate developmental test score gains of preschoolers to amount to about 3.7 IQ points per decade from 1949 to 1994 in Anglophone countries (Lynn, 2009).

However, even less accounts have provided results about further domains of childhood development, such as memory, language, or even motor development. Particularly, reports about developmental test score changes in more recent years are scarce and those that are available show inconsistent results. Whilst some studies reported increases in certain areas such as motor development of school kids (Lynn, 2009; Skogan, Oerbeck, Christiansen, Lande, & Egeland, 2018), others indicated Flynn effect reversals in their performance on Piagetian tests (Shayer & Ginsburg, 2009; Shayer, Ginsburg, & Coe, 2007). Further accounts indicated gains in personal-social, hearing-speech, or eye-hand test performance over a 30 year-period in pre-1981 Great Britain (Hanson et al., 1985). Moreover, most of the available investigations focus on examinations of specific developmental subtests only whilst especially in preschoolers, more comprehensive targeted investigations of a broad range of developmental areas by means of test batteries remain unavailable.

To contribute to understanding the nature, causes, and meaning of the Flynn effect, it seems necessary to examine test score changes in children before they attend formal schooling. Such an approach allows the assessment of the between-cohorts test score developments over time without having to expect any possible influences of changes in educational settings. To this end, we investigate here changes in the test scores of preschoolers in six developmental areas, comprising thirteen subscales, of the Viennese Developmental Test (VDT; originally termed: Wiener Entwicklungstest – WET; Kastner-Koller & Deimann, 1998, 2002, 2012), a well-established developmental test battery in German language.

2. Methods

To investigate the Flynn effect on the VDT, we used two independent sets of data that represent different data quality. On the one hand, we used the three standardization samples of the initial version of the VDT and its two revisions. These data are largely representative of the Austrian and German preschooler population aged three to six and therefore represent a high-quality data source. To support the conclusions that could be drawn from these data, we, on the other hand, analyzed time trends based on opportunistically recruited preschooler samples with a cross-temporal meta-analytical design. These data may vary due to various sample characteristics, thus introducing more statistical noise compared to the standardization samples. Therefore, IQ change scores that are based on these cross-temporal analyses may provide numerical under- or overestimates of changes in IQ points. Consequently, face-value interpretation of their numerical changes may be misleading and should be avoided. However, they serve as an excellent means in corroborating the general trend of the changes from the standardization samples in terms of observed IQ gains or losses (i.e., by showing a positive or negative sign).

Of note, preschool enrollment is high in both Austria and Germany and most children (90%+) will have attended at least one year of preschooling in both countries upon entering school at age six (Destatis, 2020; Statistik Austria, 2020). However, preschool facilities such as kindergarten or nurseries are not centrally governed in either Austria or Germany, but are regulated by regional administrations. Therefore, children are wont to have experienced different educational approaches which may have varied both between and within different regions. However, there is no reason to suspect that these approaches differ systematically in any way between the two countries.

2.1. Standardization data

We obtained the standardization data of the three editions of the VDT from the test authors. All standardization data have been purposively collected by the test authors and the test publisher in different towns and villages in both countries to be as representative as possible for Austria and Germany. This is important, because different segments of the population ability distribution have been shown not to be equally affected by test score changes (e.g., lower gains at top ability levels: Rindermann & Thompson, 2013; decreasing variability of population cognitive ability over time: Pietschnig, Tran, & Voracek, 2013). All data were collected by trained VDT administrators. Data for the original publication of the VDT in 1998 were collected in 1996 (Kastner-Koller & Deimann, 1998). The standardization sample comprised 274 Austrian children that ranged in age from three to six and that were representative for the Austrian population according to age, sex, paternal profession, and urban vs. rural place of residence. In 2001, data from another 971 children ranging from three to six years in age from Germany were collected for the second edition of the VDT (Kastner-Koller & Deimann, 2002). This sample was representative for Germany in terms of age, sex, paternal profession, but not urban vs. rural place of residence. For the standardization of the third edition, data from 261 Austrian and 124 German four-to-six year-olds were collected in 2008 and 2009 (therefore, for the purpose of our analyses, we assume a data collection year of 2008.5; Kastner-Koller & Deimann, 2012). The Austrian subsample was representative for Austria according to age, sex, parental profession and place of residence, whilst the German subsample was representative for Germany according to age, sex, and parental socio-economic status, but not place of residence. Independent sample *t*-tests showed significant differences between the Austrian and German preschooler scores in the third standardization for grammar comprehension, vocabulary, and gross motor skills ($ps < 0.05$), indicating confounding influences of nationality in change score assessments in these three subtests.

In all three cohorts, boys and girls were about evenly represented, comprising 50.4%, 50.7%, and 50.1% girls, respectively. Moreover, the number of participants was roughly the same in six half-yearly age brackets (i.e., 3.00–3.49, 3.50–3.99, 4.00–4.49, 4.50–4.99, 5.00–5.49, 5.50–5.99 years) within the cohorts. Of note, in the 2008.5 cohort under-four year-olds were not assessed. This means, that only overall comparisons between 1996 and 2001 were based on data of three-to-six year-olds, whilst all other comparisons were based on four-to-six year-olds. Importantly, the items and the administration mode of the VDT have not been changed between revisions in 12 subscales. Another subscale (i.e., block design) has remained identical between the original and its first revision. Therefore, we can comprehensively assess developmental test score changes between three time-points on 12 and between two time-points on another subscale.

2.2. Cross-temporal samples

Mean VDT subscale scores of 22 independent samples (*N*s ranged from 1077 to 1251 for different subscales; ~46% girls; mean age ≈ 4.2 years) were obtained from academic master's theses that had been supervised at the University of Vienna from 2001 to 2018. ll preschoolers whose data had been collected in the course of these theses were tested by trained VDT administrators. Once again, data from 12 subscales that had not been altered in terms of item content or administration mode were investigated.

2.3. The Viennese Developmental Test (VDT; Kastner-Koller & Deimann, 1998, 2002, 2012)

The VDT is a developmental test battery in German language that can be used with children from the ages of three to six years (excepting the coloured matrices subscale, which should be only administered starting from age four, according to the most recent revision of the VDT). The test

is administered individually and takes about 90 min in any age group and is normed for ages 3 to 5.99 years in 0.5-year intervals. The VDT is based on an interactional developmental model, which assumes that development manifests itself through the acquisition of competences that results from an interaction between the individual and the environment (Bronfenbrenner, 1981). Mainly, it is used in the context of treatment-oriented assessment of potential developmental deficits as it provides information about important domains of child development. Developmental deficits often represent a major barrier to successful schooling and school enrollment. Consequently, the VDT has conceptually been designed to assess areas that have been empirically shown to predict school success. Six such areas are assessed in the course of the VDT by means of thirteen subscales (fourteen, in the third edition) as well as a parental questionnaire.

It has long been established, that cognitive abilities are among the most important predictors of school success (e.g., Arbuckle & Mc Kinnon, 1988; or refer to Petermann, 2006, for a more recent account). Therefore, cognitive development is tested by means of four subscales (analogies, block design, coloured matrices, quiz). Another important predictor of school success pertains to lingual skills (e.g., Daseking, Lemcke, & Petermann, 2006). Therefore, language development is assessed in two subscales of the VDT (grammar comprehension, vocabulary). Memory has been shown to play an important role as well which is, for instance, reflected by positive associations between preschoolers' short-term and working memory with math achievement in school (Baddeley, 1986; Bull, Espy, & Wiebe, 2008). In the VDT, memory is assessed by the two subscales digit span and object memory.

Motor, perceptual, and cognitive development have been established to be important components of child development (e.g., Piaget, 1972). In this vein, adequate motor development appears to be necessary for more complex behaviors such as throwing a ball at a target (i.e., which requires planning, coordination, and movement; e.g., Krist, Fieberg, & Wilkening, 1993). Therefore, motor development is assessed with two subscales (gross motor skills, hand skills). Visuospatial skills have been identified as another important developmental area that show positive associations with subsequent school performance (e.g., Daseking et al., 2006). The area of visual development and visual-motor skills is assessed by the two subscales drawing and visuospatial perception. Finally, psychosocial factors appear to play a non-trivial role (e.g., Petermann, 2002). This area is assessed by a single subscale (emotions) as well as a parental questionnaire.

The subscales (but not the questionnaire) are briefly described below. All subscales yield scores that are based on dichotomous item responses (i.e., correct = 1 vs. incorrect = 0) that are summed up as the main outcome, unless stated otherwise.

2.3.1. Analogies

Here, the children need to complete sentences with a word that logically complements an expression from the first part of the sentence. For instance, "The father is a man, the mother is a…" [woman].

2.3.2. Block design

This subscale is broadly similar to other block design tests as for instance in the Wechsler test batteries. Here, coloured square tiles need to be arranged as presented by the test administrator. This subscale has been extensively revised in the most recent revision of the VDT. Therefore, no cross-temporal analyses based on this subscale and only Flynn effect assessments between the first and second, but not the third cohort are available.

2.3.3. Coloured matrices

In this subscale, a missing element out of five that completes a systematic pattern in a 3-by-3 matrix needs to be indicated, similar to Ravens-typed tests.

2.3.4. Quiz

Here, the children need to respond to questions that relate to everyday life experiences; for instance, by explaining, why it is forbidden to play on a street.

2.3.5. Grammar comprehension

The children are given a number of dolls and then read certain behaviors (e.g., "The mother allows the girl to lie down"), that they are supposed to act out with these dolls. The ensuing behavior indicates if the children were able to correctly identify the grammatical relations in the text that they had been read.

2.3.6. Vocabulary

Here, a word is being read, which the child then needs to explain. Depending on the degree of elaboration of a correct explanation, either one or two points are awarded for each word.

2.3.7. Digit span

This is a typical digit span (forward only) task in which a sequence of digits has to be repeated after it has been read to the children. Over the course of the task, the number of digits that are read increases from two to six. If a child is unable to correctly repeat a sequence of a certain length, a second sequence of the same length is read. If the child is once more unable to correctly repeat the sequence, no further sequences are presented. Two indices are used to interpret this subscale, namely (i) the number of responses (i.e., correct and incorrect) that have been recorded (i.e., this is essentially a function of how many attempts a child in a given age group needs to arrive at a correct solution for an item that is not too difficult for her) and (ii) the maximum number of digits that have been correctly reproduced.

2.3.8. Object memory

In this test, the children need to memorize the location of six objects that are located in a box with twenty drawers. Three parameters are recorded. First, the number of objects is assessed whose location is correctly recalled immediately after they have been revealed for a single time (i.e., immediate recall). Second, the number of trials that are needed to recall the location of all objects is recorded (number of trials). Third, the number of correctly recalled locations after a delay of twenty minutes is scored (delayed recall; in between the presentation and recall, 2 or 3 further subscales are administered).

2.3.9. Gross motor skills

Here, children have to complete various types of physical exercises including catching balls, balancing on one foot, or doing jumping jacks.

2.3.10. Hand skills

In this subscale, fine motor skills are assessed. The children are presented with a teddy bear that they are supposed to dress. To this end, they need to fasten a push button, a belt buckle, a knot, and a bow.

2.3.11. Drawing

The children need to reproduce a number of simple drawings from a template. Their reproductions are then rated for their correctness by the test administrator based on a number of well-defined error categories. Drawings with less than three errors receive a point, all others do not.

2.3.12. Visuospatial perception

In this task, drawings of everyday objects that have been arranged in a certain way on a card need to be matched with one that shows an identical spatial relation of these objects out of six pictures.

2.3.13. Emotions

This subscale contains pictures of persons that show unequivocal facial expressions of either happiness, anger, sadness, or show a neutral facial expression. Children are supposed to indicate the respective

displayed emotion.

2.3.14. Psychometric properties of the VDT

By means of probabilistic models (i.e., dichotomous Rasch, Partial Credit, and Mokken analyses), the subscales of the VDT have been shown to be unidimensional. Moreover, classical test theoretical analyses of standardization sample reliabilities support these favorable results (e.g., split-half reliabilities of the VDT subtests ranged from 0.77 to 0.91 in the original standardization). Orthogonal factor analytic examinations broadly support the six-factor structure of the VDT (Kastner-Koller & Deimann, 2012). Subsequent studies further support the factorial structure as well as the good discriminant, predictive (e.g., Krampen, Becker, Becker, & Thiel, 2008), and concurrent validity of the VDT (e.g., with the K-ABC; Hirschmann, Kastner-Koller, & Deimann, 2008; with the German WISC-IV: correlations of preschooler *C*-values with WISC-IV IQs = 0.77; Petermann & Petermann, 2007; and 0.78; Kastner-Koller et al., 2013).

2.4. Statistical analyses

To investigate changes between standardization samples, we first calculated pairwise standardized mean differences (Cohen *d*) between the subscale rawscores of the original and the two restandardization samples. Presently, we interpret these effect sizes according to the well-established thresholds by Cohen (1988) into $d = 0.2$ = small, $d = 0.5$ = moderate, and $d = 0.8$ = large effects. No participants under four years of age were assessed in the 2008.5 restandardization. Therefore, overall comparisons between 2008.5 and its prior standardizations are based on four-to-six year old participants only. If desired, the provided Cohen *d* values can be transformed to the IQ scale by multiplying them by 15, because Cohen *d*s can be interpreted as units of standard deviation.

For our cross-temporal meta-analyses, we applied robust modelling using the M-estimators as proposed by Venables and Ripley (2002), to minimize influences of outliers. We first predicted raw scores of each subscale by mean sample ages in linear weighted meta-regressions, to account for different participant ages. Subsequently, we predicted the resulting *z*-standardized residual values of this first set of regressions by the data collection year of the individual samples.

z-standardizing the dependent variable only allowed us to interpret the resulting linear regression slope as the amount of ability change in units of standard deviations for each year that had elapsed. Analogously to our above approach, multiplying this value by 15 transforms the slope into the IQ scale (i.e., representing annual IQ changes). We ran all analyses first for all available samples (numerical results omitted) but excluded samples with developmental deficits in another turn, to obtain test score changes based on developmentally unselected participants.

2.4.1. Measurement invariance

To corroborate the results of these analyses and ensure meaningfulness of the observed changes, we investigated IQ test score changes of the standardization samples based on latent means (we were unable to conduct measurement invariance analyses for cross-temporal data because item-level data were unavailable). Measurement invariance was established through calculating multiple group confirmatory factor analyses (MGCFA). Model fit of invariance models was evaluated according to common criteria and deemed to fit reasonably well if RMSEA was <0.08 and CFI and TLI were > 0.95 (e.g., Hu & Bentler, 1999; Kline, 2016). We followed the approach of Wu and Estabrook (2016) for binary data to restrict parameters in our invariance models.

Measurement invariance analyses for categorical data in general and binary data in particular create special circumstances compared to continuous data. Threshold parameters need to be introduced when ordered categorical variables are examined. Moreover, restricting single sets of parameters at a time is not appropriate in many instances, because identification conditions in the configural model become binding when certain constraints have been introduced. As a consequence, some restrictions cannot be examined alone compared to the configural model. In this vein, Wu and Estabrook (2016) showed, that the baseline model for invariance analyses for binary data is statistically equivalent to a model with restricted threshold parameters and loadings. This means that configural invariance should be established based on a model with constrained parameters.

Consequently, we examined configural invariance by first constraining thresholds and factor loadings across groups. Subsequently, we assessed strict factorial invariance by constraining residual variances to equality. When the CFI change between models exceeded 0.01 (Cheung & Rensvold, 2002), we iteratively freed thresholds and variances (to establish partial configural invariance) or loadings (to establish partial strict invariance) of the items with the largest modification indices to establish partial invariance (summary fit statistics are provided in the online supplementary material S1). Subsequently, we estimated latent means and calculated between-standardization samples change scores in Cohen *d*s. When models failed to converge or show reasonable fit after freeing (or omitting, in cases of configural non-invariance) a third of items, no latent means were estimated. All analyses were run twice to enable latent mean-based measurement of change calculations between 1996 and 2001 standardizations based on all participants and comparisons with the 2008.5 standardization samples based on participants that were older than three years only. Measurement invariance analyses were conducted in the Open Source Software environment R (R Core Team, 2019) by means of the package lavaan (Rosseel, 2012; the R-syntax for our invariance modelling procedure is illustrated in the online supplementary material S2).

3. Results

Test performance changes yielded a rather erratic pattern of predominantly small positive as well as negative and – for the most part – non-significant changes between the 1996, 2001, and 2008.5 standardization years. However, they were broadly similar in strength and largely consistent in terms of the sign in both the raw score- and latent mean-based calculations in our standardization samples. Therefore, we will focus on interpreting results of measurement invariant items only in our description of the observed pattern, unless indicated otherwise.

In the cognitive development area, performance on both the analogies as well as the quiz subscales appeared to have been slightly but consistently decreasing over the observed time span. However, these changes were comparatively small in strength, yielding decreases that are equivalent to about −1.8 to −2.0 IQ points over 12.5 years and did not reach nominal statistical significance (numerical values are detailed in Tables 1 and 2; numerical age group-specific changes are detailed in the online supplementary material S3 and illustrated in Fig. 1). Moreover, the changes appeared to be differentiated according to the age groups. Visual inspection of change scores suggests that especially patterns of three-year-olds often showed inconsistent results, thus possibly indicating a noisier assessment of cognitive abilities, perhaps owing to a lower test reliability in this age group. For block design, only changes between 1996 and 2001 could be assessed and no overall score could be calculated. However, the observed age group-specific changes seemed

Table 1
Test score changes per decade in Cohen *d* across three areas of cognitive development in four-to-six year-olds (raw score- [IQ points per decade]/latent mean-based changes [IQ points per decade]).

	1996–2001	2001–2008.5	1996–2008.5
Verbal reasoning (Analogies)	0.01 [0.30]/ −0.04 [−1.00]	−0.09* [−1.59]/ −0.08 [−1.40]	−0.04 [−0.50]/ −0.12 [−1.33]
Inductive reasoning (Coloured matrices)	0.23* [5.65]/ 0.20* [5.00]	−0.11 [−2.00]/ −0.12 [−2.21]	0.07 [0.72]/0.07 [0.78]
Information and knowledge (Quiz)	−0.11 [−2.64]/ −0.07 [1.75]	−0.06 [−0.97]/ −0.01 [0.18]	−0.10* [−1.08]/ −0.13 [−1.44]

Table 2
Mean raw scores according to standardization years and changes of raw scores/latent means of (partially) strict invariant items in four-to-six year-olds.

	1996			2001			2008.5			Cohen *d*s		
	n	mean	*SD*	*n*	mean	*SD*	*n*	mean	*SD*	1996–2001	2001–2008.5	1996–2008.5
Cognitive development												
Analogies ($k = 15$ items)	186	10.17	2.82	627	11.81	2.53	360	11.62	2.68	0.01/−0.04	−0.09*/ −0.08	0.04/−0.12
Coloured matrices ($k = 10$ items)	186	5.53	3.27	627	6.14	3.01	383	5.79	3.03	0.23*/0.20*	−0.11/−0.12	0.07/0.07
Quiz ($k = 11$ items per age group)[a]	186	7.42	2.39	627	7.16	2.52	385	7.02	2.32	−0.11/ −0.07	−0.06/−0.01	−0.10*/ −0.13
Language development												
Grammar comprehension ($k = 13$ items)[b]	186	9.24	2.46	628	9.29	2.50	384	8.78	2.21	0.02	−0.21**	−0.11*
Vocabulary ($k = 10$ items)	184	11.97	3.74	626	11.57	3.34	359	12.70	3.29	−0.12/ −0.14	0.34/0.32***	0.12/0.17
Memory												
Digit Span: Number ($k = 2$*5 items)[c]	147	4.96	0.91	629	4.91	1.02	382	4.87	0.99	0.05	0.04	0.05
Digit span: Span ($k = 2$*5 items)[c]	147	3.77	0.85	629	3.67	0.90	382	3.57	0.89	−0.11	−0.11	−0.12*
Object memory: Immediate recall ($k = 6$ items)[b]	147	2.82	1.29	631	3.24	1.51	383	3.58	1.29	0.29**	0.23***	0.31***
Object memory: Delayed recall ($k = 6$ items)[b]	185	5.08	0.98	631	5.14	1.03	383	5.32	0.87	0.06	0.19**	0.15*
Object memory: Number of trials[c]	184	3.87	2.19	626	4.00	2.14	383	3.94	2.02	−0.06	0.03	−0.02
Motor development												
Gross motor skills ($k = 10$ items)[d]	186	8.10	1.74	628	7.32	1.74	357	7.25	1.71	−0.45***	−0.04	−0.29***
Hand skills ($k = 4$ items)[d]	186	2.78	0.98	630	2.65	0.86	384	2.71	0.65	−0.15*	0.08	−0.05
Visual development and visual-motor coordination												
Drawing ($k = 10$ items)[e]	184	6.25	1.61	624	6.19	1.62	359	6.10	1.40	−0.03/ −0.11	−0.06/−0.01	−0.06/−0.13
Visuospatial perception ($k = 24$ items)[f]	186	14.56	5.29	629	13.83	5.43	384	14.5	4.54	−0.14/ −0.17	0.13*/0.17**	−0.01/−0.01
Psychosocial development												
Emotions ($k = 9$ items)[g]	147	6.19	1.98	627	5.76	1.89	359	5.70	1.69	−0.22**	−0.04	−0.15**

(age-specific) changes in IQ per decade that are equivalents to Cohen *d* changes in the rightmost three columns can be calculated with the following formulas: (1996–2001) = ((*d* * 15)/6) *10; (2001–2008.5) = ((*d* * 15)/8.5) *10; (1996–2008.5) = ((*d* * 15)/13.5) *10.

[a] Measurement invariance analyses were calculated separately for 3-, 4-, and 5-year-olds because different items are administered at differing ages: 1, 4, and 2 items were omitted respectively due to configural non-invariance in age-groups, overall IQ changes of measurement invariant quiz items were calculated as arithmetic means of age group-based changes and thresholds were freed for 1 item in 4-year-olds to achieve invariance.

[b] No latent mean-based changes are provided because of model non-convergence after exclusion of a third of items.

[c] No measurement invariance analyses were performed due to unsuitable response formats.

[d] No measurement invariance analyses were performed due to conceptual reasons.

[e] Thresholds and variances were freed for 2 items to achieve invariance.

[f] Two items were omitted due to configural non-invariance for latent mean estimation.

[g] No latent mean-based changes are provided because of inadequate model fit (RMSEA >0.08 and CFI and TLI < 0.90 in all invariance models); * = $p < .05$; ** = $p < .01$; *** = $p < .001$.

once more indicative of decreases rather than increases, although different age groups showed once more inconsistent results.

Interestingly, coloured matrices showed a largely consistent pattern of performance changes within a certain interval, but differing signs between these intervals. Whilst changes between 1996 and 2001 showed a small positive significant effect, corresponding to a 3.0 IQ point increase, changes between 2001 and 2008.5 were negative, indicating losses of −1.5 IQ points. This means that the overall changes in the coloured matrices performance was negligible, totaling a 1.1 IQ point increase from 1996 to 2008.5 (henceforth: the longest interval).

In our cross-temporal data, the signs of the changes for analogies, coloured matrices, and quiz were consistent with the sign of the 1996 to 2008.5 change in the standardization data. Although these findings appear to corroborate our observation of decreasing analogies and quiz but increasing coloured matrices scores (top block of Table 3; Fig. 2), the coefficient strength was once again small and effects were non-significant.

Changes in the subtests in the language development area were differentiated. Whilst grammar comprehension showed significant (albeit trivial, in terms of the effect size) performance decreases in raw scores over the longest interval, non-significant gains were observed for vocabulary. Interestingly, grammar comprehension performance virtually stagnated between 1996 and 2001 but subsequently showed a significant small negative effect, whilst vocabulary showed decreases during this fist interval and subsequently showed highly significant gains. Our cross-temporal meta-analytical data showed negative signs for both subscales (second block of Table 3), thus somewhat contrasting the findings from the standardization data and indicating ambiguity in the direction of language development changes.

Changes in the memory area were largely consistent within the subtests, but differed not entirely consistent in their direction between the subtests. In the digit span subscale, the number of recorded responses remained virtually identical across all three standardizations. However, the length of the last administered sequence appeared to consistently decrease, yielding a significant (but trivial) effect over the longest interval. Cross-temporal analyses indicated non-significant losses for both indices. Of note, we interpreted increased numbers of recorded responses as indicative of decreased performance, because more items had been incorrectly responded to. Alternatively, the increases in the number of administrations could be due to a performance increase of participants because they are able to solve more difficult items, thus necessitating a larger number of item administrations. Presently, this was not the case because the average length of the last administered sequence (i.e., representing item difficulty), decreased between the standardization samples as well as in the cross-temporal data.

In contrast, for object memory both immediate and delayed recall consistently showed gains across all three standardization samples, yielding significant trivial-to-small effects for the longest interval. The number of trials that was needed to correctly recall the location of all objects showed no meaningful changes. Consistent with these observations, immediate and delayed recall showed positive whilst number of

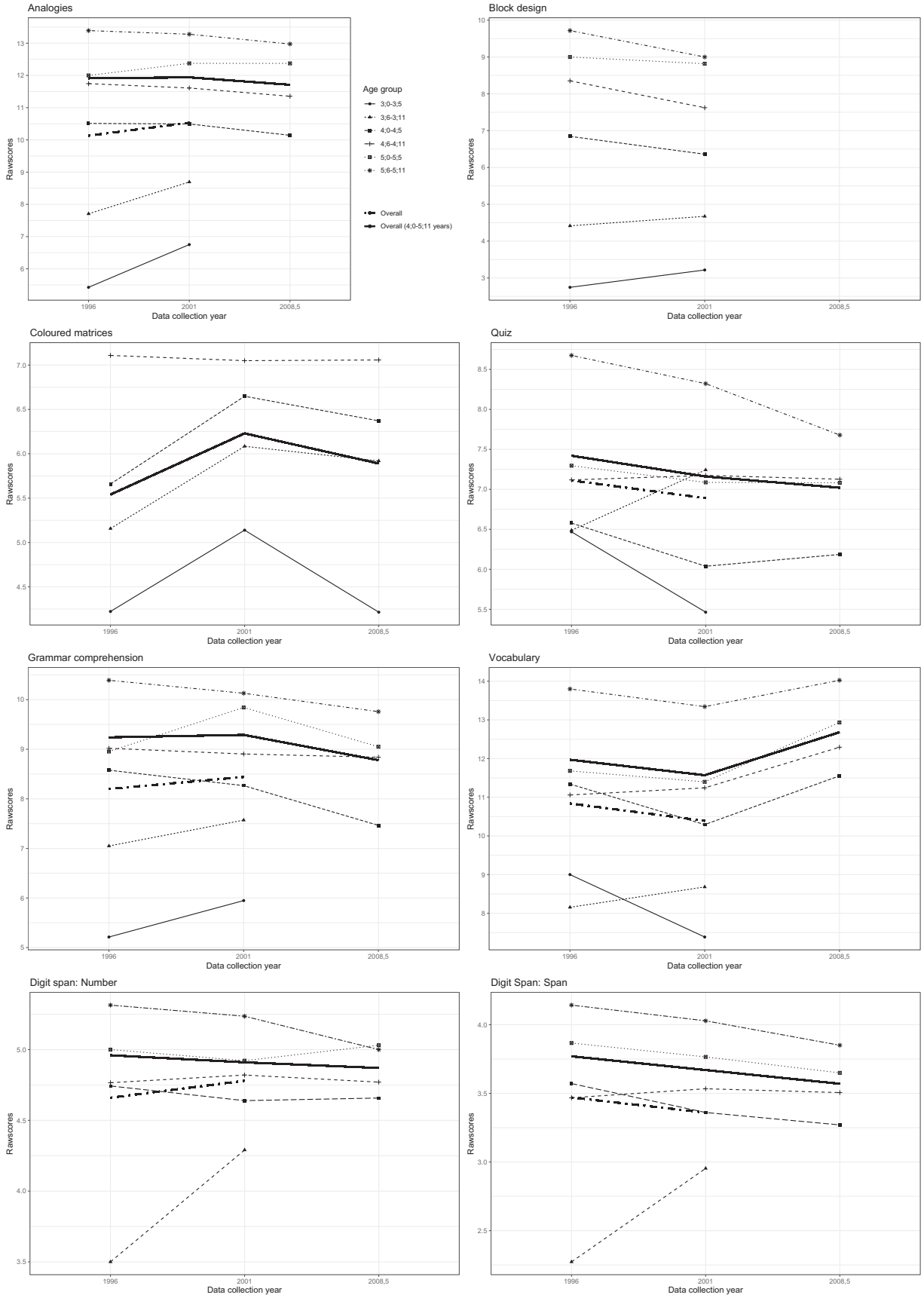

Fig. 1. Mean test score changes between data collection years of the three standardization data sets according to age group and overall.

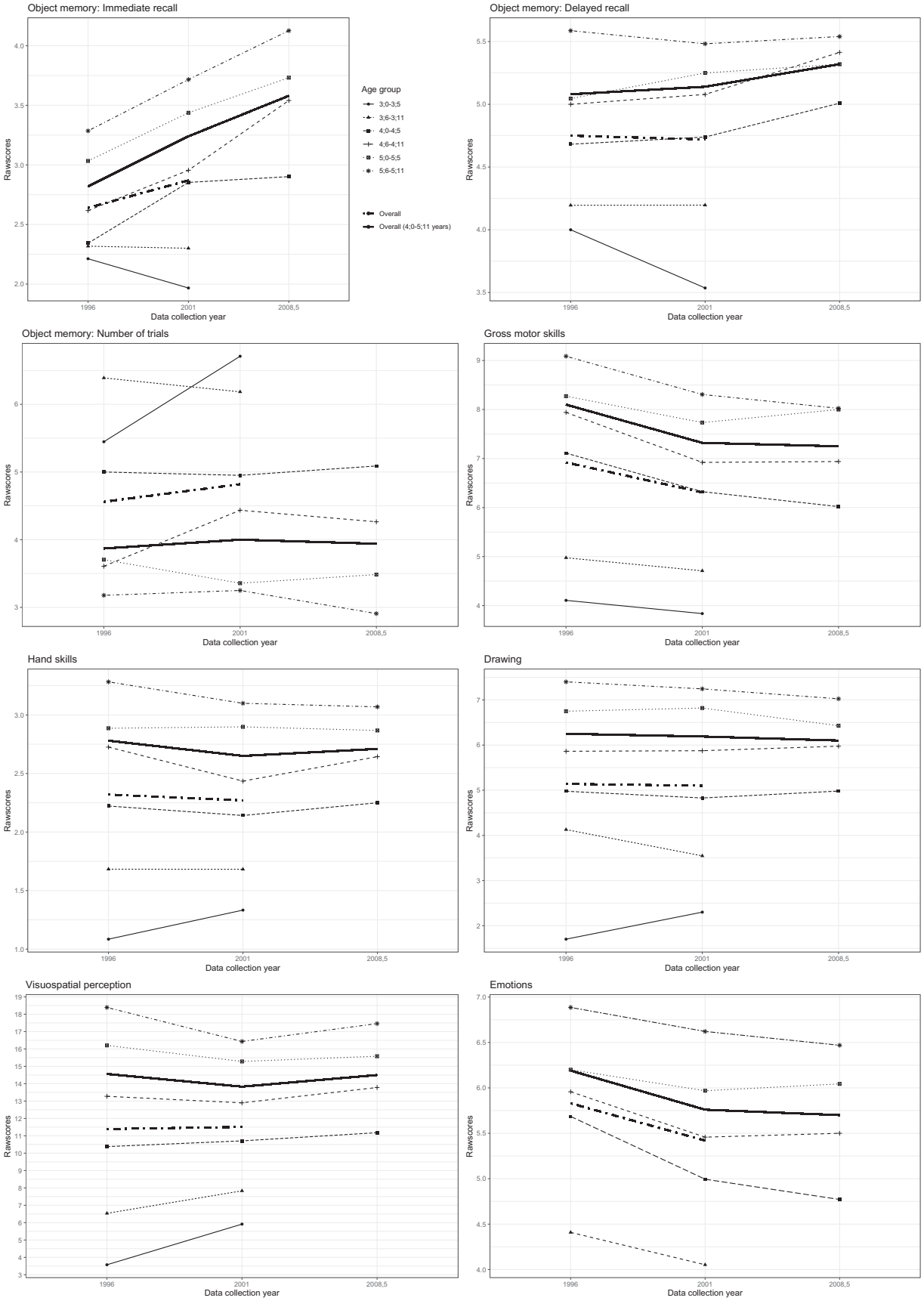

Fig. 1. (*continued*).

Table 3
Cross-temporal weighted robust meta-regressions of age-residualized subscale-scores on publication years in samples without developmental deficits, 2001–2018.

	b	*SE*	*p*
Cognitive development			
Analogies	−0.062	0.059	0.307
Coloured matrices	0.019	0.071	0.789
Quiz	−0.038	0.066	0.587
Language development			
Grammar comprehension	−0.026	0.066	0.691
Vocabulary	−0.012	0.060	0.843
Memory			
Digit span			
Number[a]	0.017	0.060	0.784
Span	−0.011	0.063	0.870
Object memory			
Immediate recall	0.085	0.068	0.228
Delayed recall	0.008	0.059	0.228
Number of trials[a]	0.046	0.044	0.295
Motor development			
Gross motor skills	−0.032	0.055	0.573
Hand skills	0.025	0.031	0.503
Visual development and visual-motor coordination			
Drawing	−0.081	0.066	0.229
Visuospatial perception	−0.076	0.036	0.052
Psychosocial development			
Emotions	0.080	0.061	0.210

$k = 16$; b = semi-standardized regression slope (i.e., indicates change in standard deviations for each year that passes); SE = standard error.
[a] Indicates that positive regression coefficients represent decreased performance.

trials showed negative non-significant changes in our cross-temporal analyses (third block of Table 3). Of note, no latent mean-based analyses were conducted for the memory area.

In the motor development area, gross motor skills showed meaningful significant negative changes over the longest interval that were driven by losses between 1996 and 2001, although changes between the second and third standardization were slightly positive. The hand skills subscale showed significant decreases between 1996 and 2001 as well, but showed increases between the second and third standardization, thus yielding virtually no change over the longest interval. The cross-temporal data corroborated the consistent negative pattern in gross motor skills, but indicated non-significant gains for hand skills (fourth block of Table 3).

For the visual development and visual-motor coordination area, we consistently observed small non-significant negative effects across all three standardization. However, for drawing initial decreases were followed by increases of about an identical strength, thus ultimately yielding no change over the longest interval. The cross-temporal analyses showed decreases for both subscales over time that barely failed to reach significance for visuospatial perception (fifth block of Table 3).

Finally, the psychosocial development area (i.e., represented by the subscale emotions only in our analyses) yielded significant trivial-to-small losses over the longest interval. The pattern was consistent across all standardization samples, although cross-temporal data showed marked changes in the opposite direction (bottom block of Table 3), thus leaving the change trajectory of psychosocial development ambiguous.

4. Discussion

4.1. Cognitive development

We show in the present study that there is little evidence for meaningful cognitive developmental test score changes in Germanophone three-to-six year-olds. The observed changes were often inconsistent in their sign according to different time spans and effect sizes were small, thus conceivably suggesting that preschoolers may have remained entirely unaffected by test score changes in the past. This would mean that the causes that have driven the Flynn effect in older samples do not become effective until school age is reached.

This surprising pattern may have not yet been noticed because accounts of test score changes on developmental tests have been limited so far and especially preschoolers have only rarely been investigated in terms of the Flynn effect. One notable exception is a review of five studies in which exclusively positive Flynn effects were observed in under three year old children from Australia, the UK, and the USA (Lynn, 2009). This contrasts our findings of mostly trivial changes whose direction was more often than not negative. However, the last cohort that had been investigated in these studies was tested in 1994, thus preceding the year in which the first cohort of our study was assessed.

Arguably, however, some common empirical observations of the Flynn effect in adults, namely strongest gains in fluid compared to smallest gains in crystallized domains (Pietschnig & Voracek, 2015) may be reflected by the present results. Specifically, the only cognitive development subscale to show an overall positive change from 1996 to 2008.5 was the coloured matrices, which represents the most fluid task of the VDT. This was supported by the positive sign of the regression slope in our cross-temporal analyses. This observation is consistent with slight improvements that have been observed for a subscale that assesses fluid domains of a Piagetian development test in British adolescents (Shayer & Ginsburg, 2009) and may be a precursor of the typical hierarchy of the domain-specificity of the Flynn effect (i.e., fluid intelligence tasks showing the strongest gains) when participants get older. However, it should be noted that the performance increased between the first and second, but decreased between the second and third standardiziation and that the change was trivial in strength, yielding a gain of a mere 1.1 IQ points from 1996 to 2008.5 (i.e., 0.78 IQ points per decade). At least in adults, fluid test score changes have been typically reported to amount to about 2 to 4 IQ points per decade, depending on the investigated timeframe (Pietschnig & Voracek, 2015).

Analogies and quiz both showed negative test score changes between 1996 and 2008.5 as well as in our cross-temporal analyses. Although the changes between standardization samples were once more trivial in size, the effects were stronger than for coloured matrices, yielding losses of −1.8 and −2.0 IQ points for analogies and quiz, respectively. Both the analogies and particularly the quiz subscale must be considered to assess more crystallized facets of cognitive abilities than the coloured matrices do. Therefore, lower gains for quiz, but also analogies (verbal reasoning domains have been shown to yield the smallest Flynn effects compared to spatial task or numerical reasoning performance in Austrian college students; Pietschnig, Voracek, & Formann, 2011) could have been expected. Our observation of negative changes, however, is surprising.

For block design, data were only available for a comparatively brief period from 1996 to 2001. However, these changes appeared to indicate negative rather than positive test score changes. Decreasing block design scores would be consistent with recent results of a meta-analysis, which indicated that spatial ability decreases may have emerged in the 1990s subsequent to a positive spatial ability Flynn effect in Germanophone adults (Pietschnig & Gittler, 2015). These results suggest that test scores of adult German-speakers have been decreasing during the time frame of our investigation. Intelligence domains in other studies that are conceptually similar to the cognitive development domains that we assessed in our study such as the verbal and spatial subscales of the General Aptitude Test Battery or the Differential Aptitude Test have yielded similar decreases in somewhat older children (Woodley & Meisenberg, 2013).

4.2. Further developmental areas

Given the well-established link of language development and intelligence (e.g., Flensborg-Madsen & Mortensen, 2019) a Flynn effect in our data seemed conceivable, in case a clear Flynn effect in cognitive test

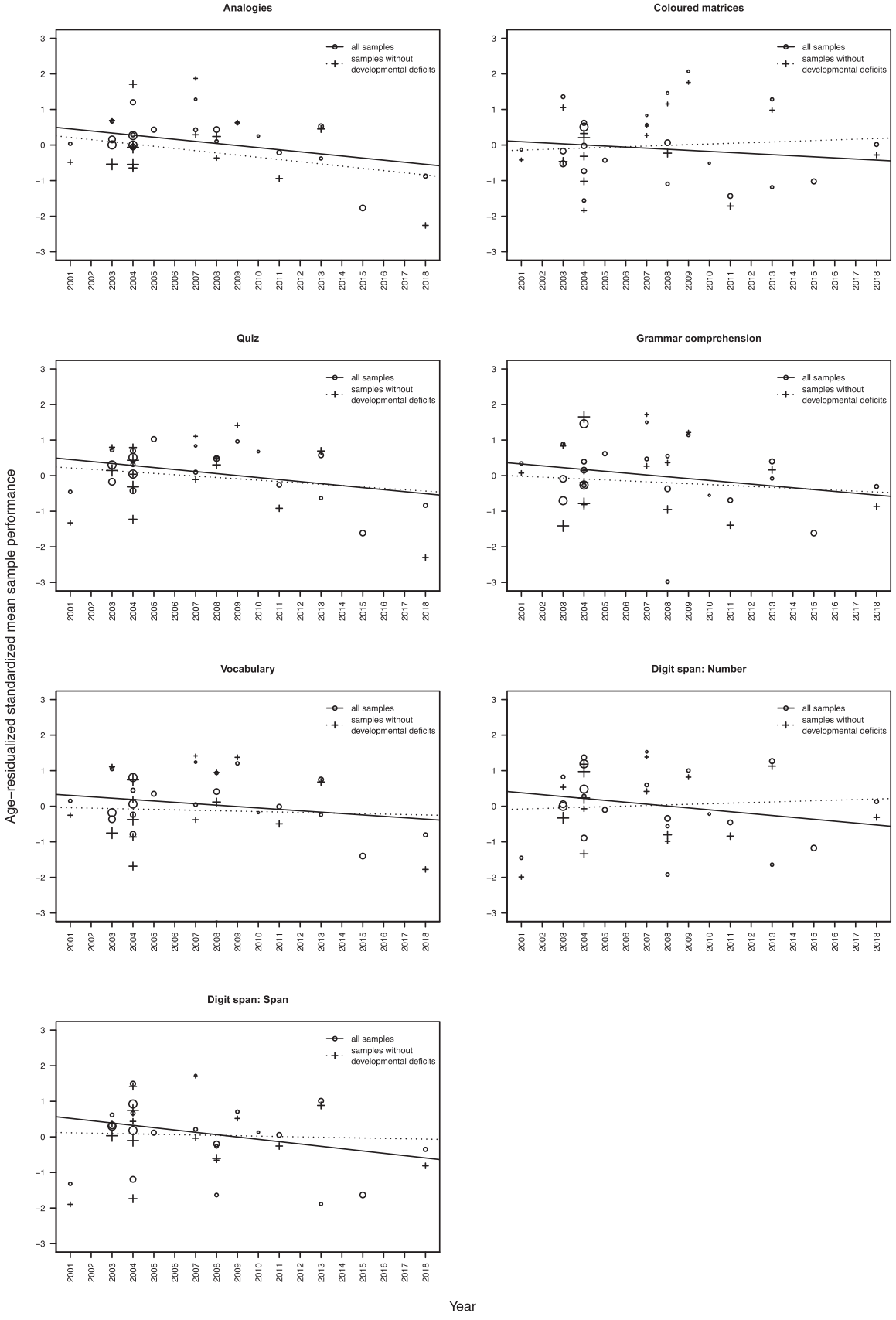

Fig. 2. Weighted cross-temporal meta-regressions of all samples and those without developmental deficits only.

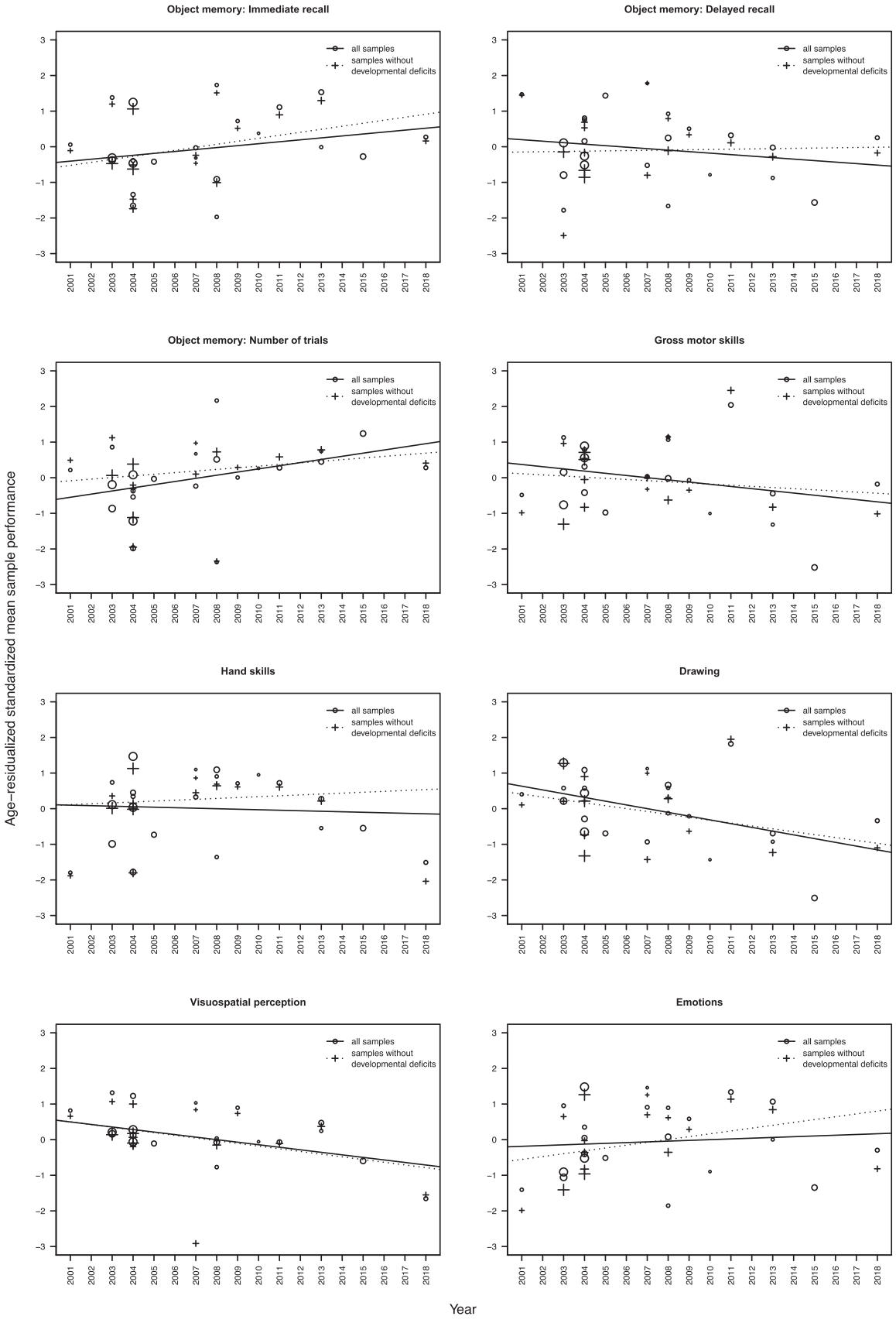

Fig. 2. (*continued*).

score changes had been observed. Interestingly, grammar comprehension showed largely consistent negative changes that emerged after the second standardization and indicated losses in both standardization data and the cross-temporal analysis. This is surprising because grammar comprehension is related to fluid abilities and therefore might be expected to show a pattern that more closely resembles the one of the coloured matrices than that of the other cognitive domains. In the light of the inconsistent change patterns in cognitive development, it may not be surprising that vocabulary showed a rather erratic pattern of gains and losses across the standardization samples and cross-temporal data.

Memory showed predominantly positive Flynn effects which were mainly driven by the object memory subscale. Whilst the results for the digit span subscale were somewhat ambiguous, object memory showed test score gains in both immediate as well as delayed recall. Over time, the preschoolers were able to recall more objects' locations correctly particularly immediately after their first demonstration but also after twenty minutes had elapsed, yielding significant small and trivial-to-small effects, respectively. The number of attempts until they had memorized the location of all objects seemed to remain fairly stable. The consistent signs in our cross-temporal analyses suggest that this pattern extends to at least 2018.

In terms of the Flynn effect, memory has been observed in the past to play a special role. Past meta-analytic accounts showed no evidence for a Flynn effect in either direction (i.e., neither positive nor negative) of several thousand adult participants from normative samples (Gignac, 2015). This fits well with our digit span findings which do not show any consistent substantial changes either. Similar inconsistent changes of digit span performances have been reported for North-American children over the last two decades of the past century (Rodgers & Wänström, 2007). In contrast to verbal or numerical memory, visual memory tasks have been observed to yield positive Flynn effects in past studies (Baxendale, 2010). The results of our object memory subscales support these findings.

In the motor development area, hand skills appeared to be largely unaffected and stable over time, gross motor development showed consistently negative changes across all standardization samples, yielding a significant small effect, as well as in our cross-temporal data up to 2018. This contrasts findings from Anglophone countries that showed increases of motor development in under-three year-olds until the mid-1990s (Lynn, 2009). However, more recent accounts indicated that Flynn effects for fine and gross motor development had considerably decelerated (Skogan et al., 2018) or ceased and reversed altogether in past decades (Shayer et al., 2007; Shayer & Ginsburg, 2009), thus conforming to our observations.

Visual development and visual-motor coordination showed little evidence for meaningful changes over time, although both drawing and visuospatial perception change scores showed consistently negative signs across the standardization samples and in the cross-temporal data up to 2018 (excepting some positive changes between the second and third standardization for perception). This contrasts prior accounts of decreasing spatial task performance in Germanophone adults (Pietschnig & Gittler, 2015).

Finally, the emotions subscale of the psychosocial development area showed significant trivial-to-small losses over the standardizations, although change scores virtually stagnated between the second and third standardization. Moreover, our cross-temporal analyses contrasted these observations, indicating test score increases up to 2018. This is most likely rooted in the differing modality of the assessed construct. From a conceptual point of view, the emotions subscale of the VDT assesses a construct that is highly similar to facets that are typically assessed in ability emotional intelligence tests. However, it has been shown that ability emotional intelligence does not show typical Flynn effect patterns, perhaps owing to the fact, that it does not reflect an ability after all, but a personality trait that is moderately correlated with some psychometric intelligence domains (Pietschnig & Gittler, 2017).

4.3. Potential causes

There are several well-established candidate theories that may contribute to explain the Flynn effect. Some that are cited often as the most likely candidate theories for the positive Flynn effect pertain to better perinatal nutrition, improved health services and hygiene, as well as longer and higher quality schooling (for an overview, see Pietschnig & Voracek, 2015). Other factors that have been specifically proposed to explain developmental test score gains relate to higher birth weights, increasing heights, as well as larger head sizes and circumferences (see Williams, 2013, for an overview), which in turn may be once again seen as a function of improvements in perinatal nutrition and health services.

All of these candidate causes are in principle suitable to explain increases or stagnation (i.e., in case that the IQ-boosting factors have hit a ceiling) of developmental test scores over time, except for schooling which cannot have any direct effect on the development of preschoolers. We did not observe any evidence for a meaningful Flynn effect in the cognitive development domains, but rather an erratic pattern of trivial changes in either direction. This may be attributed to three possible causes, all of which may be seen as equally reasonable explanations at this point.

First, the Flynn effect may only emerge after the introduction of children to formal education. This may likely be due to effects of better and longer schooling (Teasdale & Owen, 2005; Williams, 1998), but could also be attributed to delayed impacts of other candidate causes. Regardless of the responsible underlying mechanism, according to this explanation, one would expect to be able to observe positive Flynn effects in future examinations of Germanophone adults (i.e., adult cohorts that were assessed after 2008.5). The most recent accounts of Germanophone samples do not provide an entirely consistent pattern, showing evidence for both increases (Pietschnig et al., 2011) as well as decreases (Pietschnig & Gittler, 2015) in certain domains. The present failure to observe meaningful changes in cognitive development contrasts past accounts of positive developmental test score changes in Anglophone countries (Lynn, 2009). This might be interpreted as evidence for potential past developmental test score increases (i.e., prior to the first standardization data in the present paper) that may have stagnated, although the prior accounts are too sparse to rule out the possibility that preschoolers are merely unaffected by the Flynn effect.

Second, it is possible that IQ-boosting factors may have lost their potency due to ceiling effects or diminishing returns, thus having led to a stagnation of past test score gains. At least in Western countries, improvements in nutrition, health services, and hygiene may have reached a ceiling in the past decades, thus leading to a stagnation of developmental test score gains.

Finally, we cannot rule out that the presently observed stagnation may be a precursor of decreasing cognitive development test score changes that may emerge in the future and have been documented in a number of countries in adults (see, Dutton et al., 2016). Possible explanations for such anti-Flynn effects pertain to fertility (i.e., higher fertility should be associated with lower IQ scores because of selective population reproduction; Lynn, 2011), migration (i.e., lower IQ individuals migrating to countries with higher IQs, thus leading to decreasing IQs in host countries; Dutton & Lynn, 2013), or mortality (i.e., due to improvements of the modern medicine, less physically fit individuals nowadays tend to reach the reproductive age more often, thus leading to decreasing population cognitive ability because of the positive physical and psychological fitness link; Nyborg, 2012). However, direct evidence from two large independent data sets showed that neither fertility, migration, nor mortality appear to have meaningful influences on test score changes on a global scale or in Germanophone countries (Pietschnig, Voracek, & Gittler, 2018).

Another possible cause is linked to the association of test score changes with psychometric *g*. Examinations of the relation between the Flynn effect and psychometric *g* have frequently shown, that *g* is negatively associated with test score changes (Must et al., 2003; Pietschnig &

Voracek, 2015; Woodley & Meisenberg, 2013; but see Colom et al., 2001, for different findings). This may mean that past increases in specific abilities may have masked losses in psychometric *g* in the population. If this is the case, the *g*-based decreases manifest themselves, as soon as the specific ability increases have come to an end (e.g., due to ceiling effects). Consequently, a negative relation of developmental test score changes with psychometric *g* could conceivably lead to future test score decreases.

4.4. Limitations

It needs to be acknowledged, that the participants in our standardization samples came from two different nations. However, it is common for German-language ability tests to be used in all European countries with a majority of German-speakers (i.e., Austria, Germany, Liechtenstein, Switzerland) regardless of which of these countries the standardization samples originated from. For instance, the German adaptations of the Wechsler test batteries had been standardized only in Germany until as recently as 2006 (Von Aster, Neubauer, & Horn, 2006), yet they had been used in Austria, Liechtenstein, and Switzerland as well. This practice has been adopted for many tests because these countries share a common language, have similar cultural backgrounds, and have been shown to possess similar national IQ averages (Becker, 2019). Austria and Germany even share similar migration rates (World Bank, 2020) as well as Flynn effect trajectories (Pietschnig, Voracek, & Formann, 2010), thus considerably reducing potentially biasing effects of different national affiliation in our study. The results of the analyses of the standardization samples were largely in line with the findings based on the cross-temporal meta-analyses, thus corroborating the robustness of the observed results.

However, change scores in grammar comprehension, vocabulary, and gross motor skills should be taken with a grain of salt because of likely confounding influences of nationality. Moreover, it should be taken into account, that although the recruitment of the standardization sample was diligently performed in both countries, the stratification procedures differed somewhat between the time points as well as countries and are not entirely proportional to the countries' populations (i.e., in reference to the respective population, comparatively larger samples have been recruited in Austria). Therefore, potential influences of these differences cannot entirely be ruled out, although comparability of test scores between the two nationalities in all but three of the subscales in the third standardization may alleviate the most severe concerns.

We caution against interpreting the numerical strength of IQ change scores from our cross-temporal analyses at face value. The unplausibly large cross-temporal changes in several subscales may be attributed to the comparatively low number of included samples and consequently large sampling error. Still, the consistent signs of changes in many analyses of both our data sets corroborate most of our findings from our standardization data.

Finally, it needs to be acknowledged, that the change trajectories within the individual subscales were not entirely consistent across age groups. Particularly, changes in three-year-olds appeared to often yield rather inconsistent results. Consequently, we provided overall change trajectories for our analyses of standardization samples with and without inclusion of the three-year-olds to illustrate potential influences of uncharacteristic values of this age group.

5. Final words

In the present study, we showed that there is no convincing evidence for a Flynn effect in cognitive development in three-to-six year-olds on a popular developmental test battery for Germanophone preschoolers. These findings support the role of education as an important driver of test score gains. Future research needs to determine if such a pattern may be a precursor of a Flynn effect stagnation or even reversal in the adult population.

Cross-temporal meta-analysis references

*Auer, K. (2004). *Auswirkungen von Computerlernprogrammen und Computerspielen auf die visuelle Differenzierungsfähigkeit von Vorschulkindern*. Unpublished master's thesis, University of Vienna, Austria.

*Berggold, C. (2008). *Validierung des Wiener Entwicklungstests anhand des Konzentrations-Handlungsverfahrens für Vorschulkinder*. Unpublished master's thesis, University of Vienna, Austria.

*Brachner, K. (2009). *Auswirkungen der Zufriedenheit berufstätiger Mütter auf die sozio-emotionale Entwicklung ihrer Kinder*. Unpublished master's thesis, University of Vienna, Austria.

*Bruckner, J. (2004). *Händigkeit und visuelle Wahrnehmung: Ein Vergleich von links- und rechtshändigen Kindern im Alter von 4;0 bis 6;5 Jahren in Bezug auf ihre visuelle Wahrnehmung*. Unpublished master's thesis, University of Vienna, Austria.

*Büsel, M. S. (2013). *Evaluierung des Entwicklungseinschätzungsbogen der St. Nikolaus-Kindertagesheimstiftung*. Unpublished master's thesis, University of Vienna, Austria.

*Dintl, S. (2004). *Der Einfluss des sozialen Milieus auf die Entwicklung von Kindergartenkindern*. Unpublished master's thesis, University of Vienna, Austria.

*Fischer, B. (2001). *Fernsehgewohnheiten drei- bis sechsjähriger Kinder und Wirkung von Fernsehsendungen*. Unpublished master's thesis, University of Vienna, Austria.

*Haslinger, M. (2008). *Validierung der Subtests "Rechnen" und „Muster Legen" des Wiener Entwicklungstests (WET) an den Intelligence Development Scales (IDS)*. Unpublished master's thesis, University of Vienna, Austria.

*Hirschmann, N. (2004). *Die Entwicklung mathematischer Fähigkeiten im Kindergartenalter: Testanalyse der Subskala „Rechnen" für den Wiener Entwicklungstest (WET)*. Unpublished master's thesis, University of Vienna, Austria.

*Huber, P. (2013). *Zusammenhang zwischen mütterlicher Feinfühligkeit und kindlicher Entwicklung*. Unpublished master's thesis, University of Vienna, Austria.

*Jires, S. (2018). *Übereinstimmungsvalidierung der schriftsprachlichen Subtests des Wiener Entwicklungstests (WET) und des Würzburger Vorschultest (WVT)*. Unpublished master's thesis, University of Vienna, Austria.

*Kainz, S. (2003). *Die allgemeine Entwicklung von Vorschulkindern mit sozialen Problemen im Elternurteil*. Unpublished master's thesis, University of Vienna, Austria.

*Klotz, B. (2008). *Auswirkungen eines Sprachförderprogramms auf die Entwicklung von Kindergartenkindern mit Sprachentwicklungsbeeinträchtigungen*. Unpublished master's thesis, University of Vienna, Austria.

*Koch, H. (2007). *Die Entwicklung von Kindergartenkindern ihrer Kindergartenpädagoginnen und Mütter*. Unpublished master's thesis, University of Vienna, Austria.

*Lippert, M. (2015). *Entwicklungsauffälligkeiten bei Kindern im Vorschulalter: Ein Vergleich von Kindern mit deutscher und nicht-deutscher Erstsprache*. Unpublished master's thesis, University of Vienna, Austria.

*Makovec, C. M.-L. (2004). *Mütterliche Berufstätigkeit und die Entwicklung von Kindergartenkindern: Eine empirische Studie*. Unpublished master's thesis, University of Vienna, Austria.

*Mondl, M. (2004). *Einfluss der Familienform auf Entwicklung und Verhalten von Kindergartenkindern*. Unpublished master's thesis, University of Vienna, Austria.

*Montana, J. (2014). *Elterninformationen als Indikatoren für Entwicklungs- und Verhaltensauffälligkeiten*. Unpublished master's thesis, University of Vienna, Austria.

*Pabst, S. (2005). *Der Entwicklungsstand von Vorschulkindern mit und ohne Entwicklungsauffälligkeiten um Urteil ihrer Mütter: Spezielle Berücksichtigung von ratsuchenden und nicht-ratsuchenden Müttern*.

Unpublished master's thesis, University of Vienna, Austria.

*Pokorny, U. (2011). *Betreuungsformen von Kleinkindern und ihre Auswirkungen auf die Entwicklung: Haben Krippenkinder einen Entwicklungsvorsprung im Kindergarten?* Unpublished master's thesis, University of Vienna, Austria.

*Schimpl, C. (2007). *Testbarkeit von 3- bis 4-jährigen Kindern in Abhängigkeit von der Entwicklung einer Theory of Mind (ToM)*. Unpublished master's thesis, University of Vienna, Austria.

*Tratsch, M. (2003). *Konstruktion des Subtests "Rechnen" zum Wiener Entwicklungstest (WET) für 2 bis 6-jährige Kinder*. Unpublished master's thesis, University of Vienna, Austria.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.intell.2021.101544.

References[1]

Arbuckle, B. S., & Mc Kinnon, C. E. (1988). A conceptual model of determinants of children's academic achievement. *Child Study Journal, 18*, 121–147.

Baddeley, A. D. (1986). *Working memory*. Oxford, UK: Oxford University Press.

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

Becker, D. (2019). *The NIQ-dataset (V1.3.2)*. Germany: Chemnitz.

Bronfenbrenner, U. (1981). *Die Ökologie der menschlichen Entwicklung [The ecology of human development]*. Stuttgart, Germany: Klett.

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental Neuropsychology, 33*, 205–228.

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

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

Colom, R., Juan-Espinosa, M., & Garcia, L. F. (2001). The secular increase in test scores is a "Jensen effect". *Personality and Individual Differences, 30*, 553–559.

Daseking, M., Lemcke, J., & Petermann, F. (2006). Vorläuferstörungen schulischer Fertigkeiten: Erfassung von kognitiven Leistungen im Kindergartenalter [Predecessor disorders of school skills: Assessment of cognitive performance in kindergarten ages]. In U. Petermann, & F. Petermann (Eds.), *Diagnostik sonderpädagogischen Förderbedarfs [Assessement of special educational need for support]* (pp. 211–237). Göttingen, Germany: Hogrefe.

Destatis. (2020). Betreuungsquote von Kindern unter 6 Jahren nach Bundesländern. Retrieved from: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Soziales/Kindertagesbetreuung/Tabellen/betreuungsquote-2018.html.

Dutton, E., & Lynn, R. (2013). A negative Flynn effect in Finland, 1997-2009. *Intelligence, 41*, 817–820.

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

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

Flensborg-Madsen, T., & Mortensen, E. L. (2019). Language development and intelligence in midlife. *British Journal of Developmental Psychology, 37*, 269–283.

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

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171–191.

Genovese, J. E. C. (2018). Evidence of a Flynn effect in children's human figure drawings (1902-1968). *The Journal of Genetic Psychology, 179*, 176–182.

Gignac, G. E. (2015). The magical number 7 and 4 are resistant to the Flynn effect: No evidence for increases in forward or backward recall across 85 years of data. *Intelligence, 48*, 85–95.

Hanson, R., Smith, J. A., & Hume, W. (1985). Achievements of infants on items of the Griffiths scales: 1980 compared with 1950. *Child: Care, Health and Development, 11*, 91–104.

Hirschmann, N., Kastner-Koller, U., & Deimann, P. (2008). Entwicklung und Diagnostik mathematischer Fähigkeiten in der frühen Kindheit [Development and diagnosis of mathematical abilities in early childhood]. *Empirische Pädagogik, 22*, 178–192.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1–55.

Kastner-Koller, U., & Deimann, P. (1998). *Wiener Entwicklungstest (WET)*. Göttingen: Hogrefe.

Kastner-Koller, U., & Deimann, P. (2002). *Wiener Entwicklungstest (WET) (2. überarbeitete und neu normierte Aufl.)*. Göttingen: Hogrefe.

Kastner-Koller, U., & Deimann, P. (2012). *Wiener Entwicklungstest (WET) (3. überarbeitete und erweiterte Aufl.)*. Göttingen: Hogrefe.

Kastner-Koller, U., Deimann, P., Antolovic, A., Heiss, C., Kubinger, K. D., & Neumann, G. (2013). Zur Vorhersage von kognitiven Leistungen im Vorschul- und Grundschulalter: Zwei Studien zur prognostischen Validität des Wiener Entwicklungstests [Predicting cognitive performance in preschool and elementary school ages: Two studies about the predicitive validity of the Viennese Developmental Test]. *Diagnostica, 59*, 202–214.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford.

Krampen, G., Becker, M., Becker, T., & Thiel, A. C. (2008). Zur Reliabilität und Validität des "Wiener Entwicklungstests" (WET): Befunde aus drei erweiterten Replikationsstudien und Vorschläge für eine erweiterte Testauswertung [Reliability and validity of the "Viennese Developmental Test" (VDT): Results from three extended replication studies and suggestions for an extended test appraisal]. *Frühförderung Interdisziplinär, 27*, 11–23.

Krist, H., Fieberg, E. F., & Wilkening, F. (1993). Intuitive physics in action and judgement: The development of knowledge about projectile motion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1–15.

Lynn, R. (2009). What has caused the Flynn effect? Secular increases in the development quotients of infants. *Intelligence, 37*, 16–24.

Lynn, R. (2011). *Dysgenics: Genetic deterioration in modern populations* (2nd ed.). Ulster, UK: Ulster Institute for Social Research.

Must, O., Must, A., & Raudik, V. (2003). The secular rise in IQs: In Estonia, the Flynn effect is not a Jensen effect. *Intelligence, 31*, 461–471.

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

O'Keefe, P., & Rodgers, J. L. (2017). Double decomposition of level-1 variables in multilevel models: An analysis of the Flynn effect in the NSLY data. *Multivariate Behavioral Research, 52*, 630–647.

Petermann, F. (2002). Klinische Kinderpsychologie: Das Konzept der sozialen Kompetenz [Clinical child psychology: The concept of social competence]. *Zeitschrift für Psychologie, 210*, 175–185.

Petermann, F. (2006). Intelligenzdiagnostik. *Kindheit und Entwicklung, 15*, 71–75.

Petermann, F., & Petermann, U. (Eds.). (2007). *HAWIK-IV Hamburg-Wechsler-Intelligenztest für Kinder – IV*. Bern, Switzerland: Huber.

Piaget, J. (1972). *Sprechen und Denken des Kindes [Speech and thought of children]*. Düsseldorf, Germany: Schwann.

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. *Intelligence, 53*, 145–153.

Pietschnig, J., & Gittler, G. (2017). Is ability-based emotional intelligence impervious to the Flynn effect? A cross-temporal meta-analysis (2001-2015). *Intelligence, 61*, 37–45.

Pietschnig, J., Tran, U. S., & Voracek, M. (2013). Item-response theory modelling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypothesis perhaps. *Intelligence, 41*, 791–801.

Pietschnig, J., & Voracek, M. (2015). One century of global IQ gains: A formal meta-analysis of the Flynn effect (1909-2013). *Perspectives on Psychological Science, 10*, 282–306.

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

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

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

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

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

Rodgers, J. L., & Wänström, L. (2007). Identification of a Flynn effect in the NLSY: Moving from the center to the boundaries. *Intelligence, 35*, 187–196.

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

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

Shayer, M., & Ginsburg, D. (2009). Thirty years on – A large anti-Flynn effect? (II): 13- and 14-year-olds. Piagetian tests of formal operational norms 1976-2006/7. *British Journal of Educational Psychology, 79*, 409–418.

Shayer, M., Ginsburg, D., & Coe, R. (2007). Thirty years on – A large anti-Flynn effect? The Piagetian test Volume & Heaviness norms 1975-2003. *British Journal of Educational Psychology, 79*, 409–418.

Skogan, A. H., Oerbeck, B., Christiansen, C., Lande, H. L., & Egeland, J. (2018). Updated developmental norms for fine motor functions as measured by finger tapping speed and the Grooved Pegboard Test. *Developmental Neuropsychology, 43*, 551–565.

Statistik Austria. (2020). Kinderbetreuungsquoten der 0- bis 2-jährigen und 3- bis 5-jährigen Kinder 1995–2019. Retrieved from https://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/bildung/kindertagesheime_kinderbetreuung/021659.html.

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

Trahan, L., Stuebing, K. K., Hiscock, M. K., & Fletcher, J. M. (2014). The Flynn effect: A meta-analysis. *Psychological Bulletin, 140*, 1332–1360.

[1] References that are preceded by an asterisk indicate studies that were included in the cross-temporal meta-analysis.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.
Von Aster, M., Neubauer, A., & Horn, R. (2006). *Wechsler Intelligenztest für Erwachsene (WIE)*. Manual. Hardcourt Test Services.
Williams, R. L. (2013). Overview of the Flynn effect. *Intelligence, 41*, 753–764.
Williams, W. M. (1998). Are we raising smarter children today? School- and home-related influences on IQ. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 125–154). Washington, DC: American Psychological Association.
Woodley, M., & Meisenberg, G. (2013). In the Netherlands the anti-Flynn effect is a Jensen effect. *Personality and Individual Differences, 54*, 871–876.
World Bank. (2020). World development indicators: International migrant stock % of population. Retrieved from: https://data.worldbank.org/indicator/SM.POP.TOTL.ZS?locations=AT-DE.
Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika, 81*, 1014–1045.