European Journal of Psychology of Education
2003, Vol. XVIII, n° 4, 405-424

Long-term comparison (1921-2001) of numerical knowledge in three to five-and-a-half year-old children

Christine Bocéréan
Laboratoire de Psychologie (GRAPCO), Université Nancy 2, France

Jean-Paul Fischer
Laboratoire de Psychologie (GRAPCO), IUFM de Lorraine, France

André Flieller
Laboratoire de Psychologie (GRAPCO), Université Nancy 2, France

This study contributes to the debate about the Flynn effect by proposing a long-term comparison (1921-2001) of the numerical knowledge of two cohorts of three- to five-and-a-half year-old children. In 1921, Beckmann (1923) assessed the numerical development of children using four tasks (Production, Distinction, Recognition, and Naming). In 2001, we used these same tasks to test 400 children equally divided into five age groups spaced six months apart. The main results are as follows: (1) the order of difficulty of the four tasks was the same in 2001 and 1921; (2) the 2001 cohort significantly outperformed the 1921 cohort with an advance in numerical development ranging from six months to one year, depending on the task; (3) the superiority of the 2001 children showed up by the age of 3; (4) the magnitude of the rise in scores varied across tasks (the greatest gain was found for the Naming task); and (5) the children in the two cohorts used the same number-evaluation strategies, but the 2001 children used more mature strategies on the Naming task than did same-age 1921 children, particularly subitizing for apprehending small numbers. The rise in scores seems to correspond to a genuine gain in numerical ability, apparently promoted by parental child-raising practices (cross-generational transmission).

Introduction

In the international literature, the expression *Flynn effect* is used to refer to the rise in average cognitive test scores observed in many countries. When the same cognitive test is given at the same age and under the same conditions to two cohorts born in different years, the

more recent cohort generally obtains a higher mean score. Evidence of this phenomenon has been reported by many authors, especially the New Zealander, Flynn (e.g., 1987, 1994, 1998). Today, it is a well-established and widely-accepted effect. In a critical analysis of the Flynn effect, Rodgers (1998, p. 354) concluded that "even with a healthy dose of skepticism, the effect rises above purely methodological interpretation, and appears to have substantive import." The increase is indeed substantial, having been estimated at 3 IQ[1] points per decade, i.e., a fifth of a standard deviation. However, this average value varies considerably across countries, cohorts, the number of years between observations, the tested individuals' ages, and the tests used (for a recent review, see Flieller, 2001). Note in particular that the rise is much greater for nonverbal tasks than for verbal ones, and that it has tended to slow down in recent years (e.g., Teasdale & Owen, 2000).

The Flynn effect is a topic of interest for developmental and educational psychologists, for several reasons. Firstly, it raises the question of the validity of developmental norms, which turn out to be contingent upon the cohorts used to establish them, and for this reason, should be periodically revised. Whenever the same developmental scale is used to compare data obtained from cohorts tested several decades apart, large differences are observed in the typical age at which certain notions or skills are acquired. Secondly, this phenomenon raises the question of the role played by education and child-raising practices in the observed rise. Finally, the Flynn effect leads one to wonder about the ahistorical versus historically-situated nature of cognitive development. Most classic or contemporary theories implicitly or explicitly postulate that development is dictated by universal, atemporal laws. However, the rediscovery of Vygotsky's historico-cultural perspective has caused a growing number of authors to accord more importance to the historical context in which a child is developing (e.g., Bruner, 1996), to the constraints imposed by that context on development, to the possibilities it offers through collective cognitive tools the child can appropriate (concepts, representation techniques, etc.), and to the home and school learning conditions it provides, known to vary extensively throughout history.

In this light, the systematic comparison of cohorts affords some obvious methodological advantages. While it is very difficult to get a clear idea of cognitive development in different historical eras (Flieller, 2000), it is possible to compare the development of cohorts separated by a few decades, and to study what varied and what remained constant during that period. Given the rapid evolution of today's society, such comparisons should provide insight into the impact of the historical context on development. And they should be all the more interesting when the cohorts studied are separated by a lengthy time period. Unfortunately, the farther back we go in time, the less likely we are to find studies that meet the methodological standards required for replication. This is undoubtedly why cohort comparisons over very long periods (fifty years and more) are scarce.

Beckmann's (1923) study on the numerical knowledge of two- to six-year-old children is quite exceptional in this respect, not only regarding the evaluation procedures he used and the sample he observed, but also because of the information given by the author about the methodology and results. In his study, conducted in 1921, Beckmann assessed numerical knowledge using four tasks: (1) *Production*, where the child had to take the number of dice requested by the experimenter out of a box (the dice were blank), (2) *Distinction*, where the child had to choose which of two collections contained a given number of dice, (3) *Recognition*, in which the child had to point on a board to the square containing the number of dots specified by the experimenter, and (4) *Naming*, where the child had to say how many dots were in the square where the experimenter was pointing.

Beckmann emphasized that these four tasks were not equally difficult, in such a way that the same child or group of children could exhibit different levels of understanding of the same number. On a given task, a child was only accredited with knowledge of a number if he/she answered correctly several times on that number. Whenever success seemed accidental, the experimenter could go back and question the child again. This requirement prevented "shaky" success or chance responding, and thus guaranteed result stability. Another advantage of Beckmann's study is the sample tested, which he wanted to be large (N=465) and diversified

enough to include "children from all backgrounds" (see Beckmann, 1923, p. 18) attending a variety of educational institutions (public and private day-care centers, city and regional schools, etc.). The third advantage is the methodological information supplied by the author. Beckmann's description of the experimental material is highly detailed. He minutely described the dot boards used for the Recognition and Distinction tasks, stating the dimensions of the boards and squares, the diameter, color, and layout of the dots, the color of the background, etc. He also provided detailed specifications about how the tests were run (test order, instructions, etc.). Thus, although the study is very old, it is fully possible to reproduce it (even if the response scoring criteria could have been stated more precisely). In addition to methodological information, the article's detailed presentation of the results facilitates comparisons over time. The success rate observed on each of the four tasks is given for each number considered (2, 3, 4, and 5) and for each half-year of age. For the Production and Recognition tasks, the author even studied the children's response strategies and their distribution by age. The exceptional quality of this study has caught the eye of a number of specialists of numerical development, including Gelman (1972), who considered Beckmann's monograph to be one of the most extensive studies ever conducted on children between the ages of two and six. Clearly, its replication on recent cohorts is likely to furnish some useful information for understanding the Flynn effect, and for gaining insight into numerical development as well.

The meaning of the Flynn effect is controversial. The debate revolves mainly around whether or not the rise in performance on cognitive tests corresponds to true gains in cognitive abilities. For Flynn himself, the rise is illusory and does not represent any real increase in intelligence. But a growing number of authors (e.g., Flieller, 1999; Greenfield, 1998; Neisser, 1997; Williams, 1998) refuse to let themselves get trapped in a debate where cohorts have to be scaled on a single dimension, whether it be intelligence or cognitive development, both seen as one-dimensional. These authors advocate not only making cross-cohort comparisons based on a variety of tests, but also, if necessary, going right down to the item level. In line with this, Anastasi and Urbina (1997), for example, recommended using tasks with specific, well-identified content and processes. The tasks utilized by Beckmann meet this requirement, especially the Production task, where according to Gelman (1972, p. 123), children are unlikely to succeed without "responding to the numerosity *per se*".

Furthermore, a long-term replication of Beckmann's study should contribute to improving our knowledge of numerical development. If numerical development is relatively independent of current educational practices in the home and at school, then we should not find any substantial differences between the results obtained by Beckmann in the early 1920's to those obtained under the same observation conditions in the early 2000's. If, on the contrary, numerical development is dependent upon the child's social context (social pressure to use numbers, teaching of counting procedures, etc.), as contended by Bideaud (1999) for example, then inter-cohort differences are quite possible over such a long period. The differences may be relatively uniform and occur on various tasks and numbers, or they may be greater for certain tasks and certain numbers, due to their differing properties. For instance, while the numbers 1 to 3 can be apprehended as a whole very early on (via *subitizing*), numbers above 3 can only be initially apprehended by counting (Fischer, 1991). These different processes thus provide us with good material for observing number-dependent differences between cohorts. It would be particularly interesting to determine whether Beckmann's findings support Gelman's hypothesis of the primacy of counting over other, more perceptual ways of apprehending numbers (e.g., Gelman & Gallistel, 1978), and whether this same phenomenon exists today at the beginning of the 21st century.

With this twofold goal in mind, we set out to replicate Beckmann's study and to answer the following questions: (1) Using Beckmann's numerical tests, do we find the mean rise in performance observed on other tests? (2) If so, is the rise of equal magnitude? (3) Is the rise uniform or does it vary across tasks and numbers? (4) Is the difficulty order of the four tests the same in 2001 as it was in 1921? (5) Do the children use the same strategies? 6) If there are any strategy differences, can they account for differences in performance?

A preliminary question must be raised, however: Is it legitimate to compare the results of two studies conducted in different countries, Germany for Beckmann's study and France for ours? Potential language differences (regularities, pronounceability, etc.) do not pose a problem here: the names of the numbers are one-syllable words in both languages. Moreover, in a comparison of Beckmann's study and Descoeudres's (1921) work on French-speaking Swiss children, Stern (1927) found a remarkable degree of concordance between the mean ages at which the numbers 2, 3, and 4 were mastered. It follows, then, that Beckmann would probably have obtained similar data had he conducted the study in France. Furthermore, a comparison of current German and French norms for the K-ABC arithmetic test (Kaufman & Kaufman, 1993) revealed the superiority of German children in eight of the ten three-month age groups included in our study (from 3 to 5 1/2 years), equal performance in one age group, and superior French performance in only one group. In other words, today's German children as a whole obtained higher scores than the French children did. Hence, if the hypothesis of a rise in numerical performance between 1921 and 2001 is validated in France, we can conclude, *a fortiori*, that it would also be validated in Germany.

Method

Beckmann's (1923) four tests were used. We took the utmost care to remain as close as possible to the original tests so that we could compare the performance of children in 1921 with that of children 80 years later. The present section describes the experimental paradigm in detail.

Tests: Materials, procedure, instructions, and scoring of responses and strategies

The same questioning scheme was used for the Distinction, Recognition, and Naming tests. Each test started with the number 2, and then went on to 3, 4, and 5. Whenever the child failed on a number, he/she was tested again on the number just below it. If success was achieved on the lower number, the child was considered to have acquired the lower number. For the Production test, where the child was less likely (than on the Distinction test) to succeed by chance or to succeed because of some particular arrangement of the objects (as in the Recognition test), we limited the test to two correct productions of a number.

The four tasks are described below.

Production test

Materials. The child was given a short, cylindrical metal box 17.5 cm in diameter containing 60 blank dice with 16 mm sides (Beckmann's dice measured 15 mm, but we were unable to find blank dice of that size).

Procedure and instructions. The child had to take n dice for him/herself ("Take n dice") and then give n to the experimenter ("Give me n dice"). The number $n+1$ was tested if both productions of n were successful.

Response scoring. A number was granted definitively to the child if he/she succeeded twice.

Strategy scoring. The way the dice were grasped was noted on the first production. The four possible strategies were explicit counting, simultaneous grasping, grasping by decomposition, and one-by-one grasping. The strategies are presented in detail in the "Strategies" section.

Distinction test

According to Beckmann, the distinction process could be triggered in two ways on this test. He gave the example of the distinction of 2, which could be assessed by the question "2, is it this or is it that?" or the question "Is this 2 or is it 3?"). Insofar as no further information was given, we decided to ask both questions.

Materials. The dice described above were used.

Procedure and instructions. Because the task and instructions are relatively difficult to state in full, we will simply present a specific example: the number 3. The experimenter laid out two collections in front of the child, one of 3 dice and the other of 2 dice. The child had to point to the collection that was the right answer to the question: "3, is it this or is it that?" Then, with the collection of 3 dice, the child had to answer the question: "Is this 3 or is it 4?" If the child failed on 3, the experimenter laid out two collections, one of 2 and the other of 3 dice, and the child had to point to the collection that was the right answer to the question: "2, is it this or is it that?" Then, in the presence of 2 dice, the child had to answer the question: "Is this 1 or is it 2?" The correct answers were located on the child's left first, and then on the right. The number $n+1$ was tested if success was achieved on both ways of distinguishing n (both questions).

Response scoring. We had decided in advance that success would be granted to a number with four correct answers (two per type of distinction). But in the end, we changed this requirement to two correct answers (one per type of distinction) for the following reasons: (1) For numbers above 2 or 3, the probability of random success was low: for 5, for example, it was only 1/256. (2) With four correct answers as the requirement, this would have been the only test where not a single error was tolerated.

Recognition test

Materials. In Beckmann's detailed description of the dot boards used for several Recognition and Naming tests, he stated the size of the boards, the size of the squares containing the dots, the size of the dots, and their layout in the squares (see Beckmann, 1923, p. 5 and sq.). However, we still were not sure about the boards used for the Recognition task because two types of materials were possible: (1) The first was to have one board per number tested. In this case, each board contained squares showing the number of dots in question, plus squares containing smaller numbers (e.g., the board for 3 contained squares with three dots, two dots, and one dot). The dots were laid out in different ways whenever possible (in a line, forming a corner, randomly). (2) The second possibility was to have a single board (the board for 5) with squares containing one to six dots (two squares per number). We constructed two boards of this type (see Figure 1); on the first, the dots either formed a line or were in domino configurations, and on the second, they formed corners or were randomly arranged.

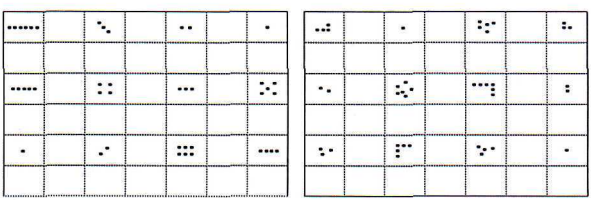


Figure 1. Dot boards used in the recognition and naming tests (actual dimensions: 21x16cm)

In a pre-experiment on 44 three- to six-year-old children, we used the first type of material. The test turned out to be abnormally easy, since a child could answer correctly by default: a given number (say 3) could be "recognized" even when only smaller numbers were mastered (1 and 2). So for the actual experiment, we used the second type of material. However, since the problem of what material Beckmann had used was not fully clear, several rereadings of his

paper led us to question our choice, and we became convinced that it did not in fact correspond to the original experiment. It therefore seemed preferable to estimate the scores we would have obtained with the first kind of material, using the method described below, rather than to base the results obtained by our children on materials that were not comparable to Beckmann's.

Procedure and instructions. The child had to point to a collection of *n* dots on the first board ("Show me *n*"), and then a second collection of the same number ("Show me another *n*"). Children who failed on *n* were asked to point to *n-1* on the second board. The number *n+1* was tested if the child found both squares containing *n* dots on the first board.

Response scoring. The score *n* was assigned to children who succeeded four times (twice on each board). However, one recognition error was tolerated if it was corrected by the child.

Score estimation. The recognition scores were estimated from the pre-experiment data for the 37 children who were between three- and five-and-a-half years old. Two arguments justify this estimation procedure: (1) There was no significant difference between the pre-experiment and the experiment in the means obtained on the Production, Distinction, and Naming tests, from which the recognition scores were estimated ($F(1,435)=2.68$, $p=.10$ for production, $F(1,435)=1.71$, $p=.19$ for distinction, and $F(1,435)=1.34$, $p=.25$ for naming). (2) The pairwise correlations among the pre-experiment tests were not different from the pairwise correlations among the experiment tests (*z*-values: 1.19 for the production-naming pair, 1.74 for the production-distinction pair, and 1.70 for the naming-distinction pair, $p>.05$).

A multiple regression analysis (without a constant term[2]) was conducted on the pre-experiment data, with the recognition score as the dependent variable and the production, distinction, and naming scores as explanatory variables. Given that the fit was very good ($R2=.98$, $F(3,34)=532.3$, $p<.001$), the predicted data should be very reliable. Based on the regression equation, predicted recognition scores were calculated for the 400 children, for comparison with Beckmann's subjects.

Naming test

Materials. The dot boards shown in Figure 1 were also used for the Naming test. Beckmann did not state which board or boards of *n* dots he used for this test, but this was not a problem because, unlike recognition, naming a number of dots in a given square could hardly be influenced by the numbers represented in the other squares.

Procedure and instructions. The experimenter pointed to the two collections of *n* dots on the first board, and asked "How many dots are there here?" If the child made a mistake, he/she was prompted to reconsider and try again. The number *n+1* was tested if the child answered correctly twice for the number *n* twice. If the child failed on *n*, he/she was asked to name *n-1* on the second board.

Response scoring. Children who made a mistake were prompted to think and try again. The score *n* was assigned when the child succeeded four times (twice on each board). Mastery of the cardinality principle (Gelman & Gallistel, 1978) was required. For example, children who settled for counting "1, 2, 3" were asked again "How many dots are there here?" and were expected to say "3".

Strategy scoring. The experimenter noted the strategy used by the child to name the lined-up dot collections. The possible strategies were direct naming, inferred counting, and explicit counting. For details, see the "Strategies" section below.

Sample

Beckmann tested children between the ages of two and six-and-a-half. Our pre-experiment pointed out what range of ages we would use in our sample: children under three

(whose behavior lacked stability) and children over five and a half (high risk of a ceiling effect) were not included in the study.

The participants fell into five age groups (AG) each covering six months. The first age group (midpoint 3;3) ranged from three to three-and-a-half years; the last age group (midpoint 5;3), from five to five-and-a-half years. Each group contained 80 subjects. The 400 children were attending one of nine preschools in the Moselle Department of France. Only 6% of the initially-selected children could not be tested, 4% because the parents refused and 2% because the child refused.

For each age group, the girl/boy ratio was controlled, along with the socio-occupational category of the head of the household. Based on the quota method, each age group had to be representative of the French population of individuals over 18. This method ensured not only the representativeness of the sample with respect to sex and socio-occupational category, but also comparability across age groups.

Testing

The children were questioned individually in a room near their classroom. The order of the tests (Production, Distinction, Recognition, and Naming) was reversed every other time. The testing was done by two of the present authors, each of whom questioned 200 children, 40 per age group. A joint practice session was held to ensure uniform testing procedures. The data were gathered between October 2000 and May 2001.

Results

First, we checked to make sure that there was no testing-order effect (all p's were above .60) and no experimenter effect (all p's were above .15). We also made sure that, as a whole, the children in the five age groups had been tested at about the same time of year: the ANOVA yielded an effect that was marginally significant ($F(4,395)=2.39$, $p=.05$) but extremely weak ($eta2=.024$).

The raw scores from 1921 were reconstructed from the data supplied in the original article. The sizes of the age groups (3;3 to 5;3) in Beckmann's study were 46, 25, 41, 42, and 56, making for a total of 210 children.

Development curves

Figures 2a and 2d bring out the rise in mean scores by age, for each test and cohort. Analyses of variance on each 2001 test confirmed that the score differences across age groups were highly significant ($p<.001$): $F(4,395)=108.50$ for production, $F(4,395)=88.43$ for distinction, $F(4,395)=129.41$ for recognition, and $F(4,395)=90.24$ for naming. The same was true of the 1921 test scores ($p<.001$): $F(4,205)=28.15$ for production, $F(4,205)=33.47$ for distinction, $F(4,205)=33.28$ for recognition, and $F(4,205)=24.27$ for naming. It turned out, then, that Beckmann's tests clearly discriminated the age groups, so they can be deemed well-suited to assessing what they were designed to assess.

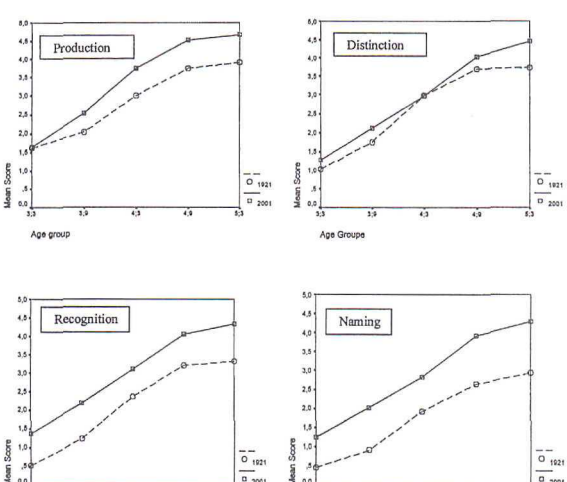
Test difficulty

Table 5 gives the mean scores of the two cohorts on each test. These figures correspond to the means of the numbers mastered on each test. Given that in 1921, the age groups were of unequal sizes, they were weighted so as to have equal weights in the analysis[3]. We can see that in 1921, the tests could be ranked as follows (in increasing order of difficulty):

Production < Distinction < Recognition < Naming

Basically the same order was found in 2001, with one exception: the Distinction and Recognition tests were of equal difficulty ($t(399)=-1.17$, $p=.24$), which gave us the following order:

Production < Distinction = Recognition < Naming



Figures 2a to 2d. Mean score on each test, by age group and cohort.

Comparison of performance on each number, by test and cohort

Beckmann reported the percentage of children in each age group who successfully produced, distinguished, recognized, and named each number tested. We calculated these figures for the 2001 cohort, except for the Recognition test (the regression did not generate whole numbers). This allowed us to compare the numerical performance of children in 1921 and 2001 on the Production, Distinction, and Naming tests (see Tables 1 to 4).

Production

As a whole, the 2001 cohort almost always outperformed the 1921 cohort (see Table 1). However, there was little difference between the scores obtained for the number 2, where every age group performed equally well as its counterpart in the other cohort. The superiority of the 2001 participants was quite marked for numbers 3 and above, and for age groups 4;3 and older. In general, the 2001 children achieved a success rate above that of 1921 children who were at least six months older, and in two cases, the 2001 group was even a year ahead (for the production of 3, the 2001 children in the 4;3 age group outperformed the 1921

children in the 5;3 age group; for the production of 5, the 2001 children in the 4;9 age group outperformed the 1921 children in the 5;9 age group of Beckmann's study). In short, the difference between the two cohorts ranged between six months and one year, in favour of the 2001 children.

Table 1

Mean score, number of subjects, and standard deviation on each test, by cohort

Experiment Year		Production	Distinction	Recognition	Naming
1921	Mean	2.86	2.64	2.13	1.78
N=201	Standard deviation	1.52	1.71	1.75	1.70
2001	Mean	3.42	2.97	3.01	2.86
N=400	Standard Deviation	1.61	1.70	1.49	1.65

Distinction

The superiority of the 2001 cohort (see Table 2) was much less pronounced on this test. The percentage of children who succeeded in distinguishing a given number was sometimes even lower in 2001 (e.g., on the numbers 4 and 5 for group 4;3). However, for the two oldest age groups, the 2001 cohort was about 6 months ahead of the 1921 cohort.

Table 2

Percentage of children who successfully produced the numbers 2, 3, 4, and 5, by cohort and age group (AG)

	AG	3;3	3;9	4;3	4;9	5;3	5;9
	*N*1921	46	25	41	42	56	60
Number Tested	*N*2001	80	80	80	80	80	
2	1921	69.6	84.0	90.2	89.9	100.0	100.0
	2001	71.4	87.6	97.6	100.0	98.8	
3	1921	19.6	20.0	63.4	83.3	82.1	93.3
	2001	16.4	51.3	83.8	92.6	98.8	
4	1921	4.4	12.0	39.0	54.8	64.3	86.7
	2001	5.1	21.3	56.3	85.1	88.8	
5	1921	0.0	4.0	17.1	35.7	44.6	70.0
	2001	1.3	6.3	38.8	73.8	81.3	

Naming

For all numbers and age groups, the 2001 cohort obtained better naming scores than the 1921 cohort (see Table 3). The 1921 cohort lagged behind by six months (AG 3;9 in 2001 were better at naming 2 than AG 4;3 in 1921; AG 4;9 in 2001 were much better at naming 4 than AG 5;3 in 1921) or even a year (AG 4;3 in 2001 named 3 better than did AG 5;3 in 1921; AG 4;9 in 2001 named 3 better than did AG 5;9 in 1921).

Synthesis

To get an overall view of the cross-cohort comparisons on these three tests, we used an external acquisition criterion commonly employed in developmental psychology: a number

was regarded as acquired by a given age group if it was mastered by at least 75% of the children in that group. Table 4 presents the age at which each number was considered acquired on the three tests in 1921 and in 2001. We can see that the numbers were acquired earlier in 2001 on every test. Moreover, the observed gap between the acquisition ages increased as the number increased. It was at its minimum (one month) for the production and distinction of 2, but reached one year and five months for the production of 5.

Table 3

Percentage of children who successfully distinguished the numbers 2, 3, 4, and 5, by cohort and age group (AG)

	AG	3;3	3;9	4;3	4;9	5;3	5;9
	*N*1921	46	25	41	42	56	60
Number Tested	*N*2001	80	80	80	80	80	
2	1921	47.8	73.0	87.8	98.9	94.6	98.3
	2001	55.1	76.4	92.6	98.8	98.8	
3	1921	6.5	16.0	63.4	81.7	78.6	91.7
	2001	15.1	38.9	65.1	91.3	92.6	
4	1921	2.2	12.0	36.6	52.4	57.1	85.0
	2001	3.8	12.6	30.1	68.8	83.8	
5	1921	0.0	4.0	21.9	35.7	39.3	73.3
	2001	0.0	8.8	16.3	45.0	71.3	

Table 4

Percentage of children who successfully named the numbers 2, 3, 4, and 5, by cohort and age

	AG	3;3	3;9	4;3	4;9	5;3	5;9
	*N*1921	46	25	41	42	56	60
Number tested	*N*2001	80	80	80	80	80	
2	1921	21.7	36.0	63.4	83.4	91.5	93.4
	2001	57.6	78.8	91.4	98.9	100.0	
3	1921	2.2	12.0	34.1	66.7	54.0	85.0
	2001	7.6	31.3	57.6	87.6	93.8	
4	1921	0.0	8.0	19.5	21.5	37.7	70.0
	2001	1.3	8.8	26.3	61.3	75.0	
5	1921	0.0	0.0	12.2	9.5	21.4	51.7
	2001	0.0	5.0	17.5	46.3	60.0	

Cohort effect on the mean score of the four tests

For each test, Figures 2a to 2d compare the mean scores of the two cohorts by age group. The curves suggest a systematic increase in scores between 1921 and 2001, with two exceptions: (1) the 3;3 age groups had equivalent mean production scores in 1921 and in 2001, and (2) the 4;3 age groups had equivalent mean distinction scores in 1921 and in 2001.

A Cohort (2) x Age Group (5) analysis of variance on the scores of each test yielded a significant main cohort effect ($F(1,610)=30.41$, $p<.001$ for production; $F(1,610)=8.73$, $p<.003$ for distinction; $F(1,610)=79.40$, $p<.001$ for recognition; $F(1,610)=94.53$, $p<.001$ for naming) and a significant main age-group effect ($F(4,610)=112.29$, $p<.001$ for production; $F(4,610)=108.51$, $p<.001$ for distinction; $F(4,610)=136.26$, $p<.001$ for recognition;

$F(4,610)=96.23$, $p<.001$ for naming). On the other hand, none of the four cohorts by age-group interactions was significant at the .05 level.

Note that the superiority of the 2001 cohort over the 1921 cohort was observed by the age of three on the Recognition ($F(1,120)=22.96$, $p<.001$) and Naming ($F(1,120)=15.41$, $p<.001$) tests, as well as on overall performance (mean of the four test scores, $F(1,120)=7.11$, $p<.01$).

The magnitude of the rise, illustrated in Figures 4a to 4d, can be quantified using Cohen's (1969) *d*, which is equal to the difference between the mean scores of the two cohorts divided by the standard deviation of the oldest cohort. The rise magnitudes obtained from the data in Table 5, in decreasing order, are $d=0.64$ for naming, $d=0.50$ for recognition, $d=0.37$ for production, and $d=0.19$ for distinction.

Table 5

Age (years; months) at which the numbers 2, 3, 4, and 5 were acquired, by test and cohort

		Number tested			
Task		2	3	4	5
Production	1921	3;5	4;6	5;6	>6;3
	2001	3;4	4;1	4;7	4;10
Distinction	1921	3;10	4;7	5;7	>6;3
	2001	3;9	4;5	5;0	>5;3
Naming	1921	4;6	5;7	>6;3	>6;3
	2001	3;8	4;6	5;3	>5;3

Given that Beckmann's sample was not described with as much precision as one would like, it is interesting to compare the cohorts by looking solely at the 2001 children from underprivileged homes (laborer or unemployed head of household). Here again, the scores of the 2001 cohort were significantly higher for production ($F(1,418)=9.35$, $p=.002$), naming ($F(1,418)=35.84$, $p<.001$), and recognition ($F(1,418)=25.11$, $p<.001$). This gave us the following rise magnitudes: $d=0.56$ for naming, $d=0.45$ for recognition, $d=0.30$ for production. These increases are a little smaller than when the analyses dealt with all 2001 participants pooled, but they mean the same thing.

Strategies

Beckmann studied the strategies used on two of his tests, Production and Naming. For 2001, two of the experimenters coded the strategies in the 400 protocols. The initial agreement rates between the two judges, measured using Cohen's kappa, were as high as .97 for production and .96 for naming. The rare disagreements were discussed and resolved.

Production of the number 3

Beckmann identified three production strategies (which he called "types") that led to success for the number 3. They correspond to what we called one-by-one grasping (Beckmann's *unity*), simultaneous grasping (*group*), and grasping by decomposition (*sum*). In the *unity* type, the child grasped one die at a time; in the *group* type, the child grasped all of the required dice in one handful; in the *sum* type, the child grasped the dice in groups of more than one (e.g., 2 dice then 3 dice for the number 5).

Beckmann's critical remark regarding the strategies was that the *unity* type was the easiest production behavior: it was used less and less often with age, while the *group* strategy became more and more prevalent. This regular shift was not observed in 2001 (see Table 6).

We noted two principal differences between our results and Beckmann's: (1) The percentage of children who used the *sum* strategy in 1921 was systematically above the 2001 figure (this cohort difference was statistically significant in a two-tailed Mann-Whitney test: $U=0$, $p<.01$). (2) The 2001 children used the *group* strategy more than the 1921 children did (this difference was marginally significant: $U=1$, $p=.057$, two-tailed test).

Table 6

Percentage of children who used each strategy to correctly produce the number 3, by age group (AG) and cohort

Type	AG	3;3+3;9	4;3	4;9	5;3	All AG's combined
Unity	1921	58.3	56.7	45.5	39.1	49.9
	2001	48.2	49.3	37.8	50.6	46.5
Sum	1921	33.3	25	30.3	30.4	29.8
	2001	11.1	19.4	13.5	21.5	16.4
Group	1921	8.3	18.3	24.2	30.5	20.3
	2001	40.7	31.3	48.7	27.9	37.2

Note. The 1921 sample size was not indicated by Beckmann; the 2001 sample contained 274 children.

To assess the effectiveness of the strategies, Beckmann examined overall performance (mean score on the four tests) as a function of the strategy used to produce 3. He concluded that "it was the representatives of the *group* type who obtained the highest performance" (Beckmann, 1923, p. 27). Our data do not allow us to draw such a conclusion. An Age Group (4) x Strategy (3) analysis of variance (with overall score as the dependent variable) did not yield a difference across strategies ($F(6,274)=.62$, $p=.72$).

According to Beckmann, then, performance is linked to the strategy used, yet our results indicated variable production modes that do not form a hierarchy. The *group* strategy which, according to Beckmann was the most advanced, clearly reveals mastery of the number, but mastery here does not imply the sole use of this strategy. The *unity* grasping mode is an infallible procedure that certain children may prefer over others. As for the *sum* strategy, there are two possible routes for arriving at 3: grasping one then two dice, or grasping two then one. This strategy may represent a true addition rather than a step-by-step approach on the part of the child, or it could even indicate the motor inability to grasp three dice at the same time. The Production test thus seems to hinge upon external factors related to the handling of the dice (taking them out of the box), to their random locations in a relatively small box, etc.

Naming

We distinguished three strategies: *direct apprehension*, *explicit counting*, and *inferred counting*. *Direct apprehension* consisted in stating the number in question without exhibiting any traces of the strategy used. In many cases it corresponded to simple visual recognition. However, for some numbers (especially high ones), this strategy could result from a highly internalized counting process or from a more sophisticated strategy memorized from arithmetic (e.g., 2+2=4). *Explicit counting* was counting aloud one number at a time, starting at 1. Inferred *counting* was counting but not aloud; it was only revealed by external signs such as pointing to the dice one by one, or eye or lip movements.

Beckmann defined two strategies, *explicit counting* and what he called *recognition*, a strategy which corresponds to what we called *direct apprehension* (to avoid confusion between the *recognition* strategy for naming, and the Recognition test, our terminology will be used here). Beckmann noted the naming strategies of 145 children. Table 7 presents the

number of children in his sample who used each type of strategy, by age group and number tested. He interpreted these results by explaining that a given number is named by counting first, and then by direct apprehension as the child grows older. Beckmann (1923, p. 31) tells us, "The number 3, for example, is recognized by counting until about 4;9, but then recognition as a complex takes over; the number 4 is preferentially apprehended by counting until 5;9, but only later (after 6 years) by recognition."

Table 7

For each number tested, percentage of 1921 children who used each naming strategy (C for counting and DA for direct apprehension), by age group (AG)

	AG	4;3	4;9	5;3	5;9	6;3
Number tested	Strategy	*N*=48	*N*=26	*N*=27	*N*=20	*N*=24
2	C	25.0	16.6	11.1	3.9	2.3
	DA	75.0	83.4	88.9	96.1	97.7
3	C	71.4	33.3	34.7	20.8	4.6
	DA	28.6	66.6	65.3	79.2	95.4
4	C	80.0	80.0	70.8	52.0	17.4
	DA	20.0	20.0	29.2	48.0	82.6
5	C	100.0	87.5	76.1	60.9	33.3
	DA	/	12.5	23.9	39.1	66.7
6	C	100.0	96.4	90.0	63.7	41.3
	DA	/	3.6	10.0	36.3	58.7

In order to draw up a table comparable to Beckmann's, we combined our explicit and inferred counting strategies into a single category (see Table 8). An analysis of the types of naming strategies used by the children in 2001, and their comparison to the 1921 strategies, pointed out three important findings:

1 A fine-grained analysis of the 2001 naming strategies (with all three initial types) indicated an impressive amount of regularity in the ordering of the types during child development. In all cases except one, i.e., for the 222 children who were at least capable of naming the number 3, the following within-individual strategy order was found for the naming of higher and higher numbers: *direct apprehension>inferred counting>explicit counting*[4]. For example, a child who succeeded on 5 via explicit counting had used inferred counting to name 4 and direct apprehension to name 3. But a child who succeeded on 5 by direct apprehension had necessarily done likewise on 4, 3, and 2. This result is consistent with Beckmann's claim that at a given age, a child directly recognizes lower numbers, but has to count on higher numbers. In general, the detailed results given in Table 8 point in this direction: as age increased, counting became less frequent and direct apprehension started to take over.

2 For the number 2, naming behavior was the same in 2001 as it was in 1921: today's children overwhelmingly apprehended this number directly by the age of three.

3 In contrast, the strategy used for the number 3 differed: 71.4% of Beckmann's four to four-and-a-half year-olds used counting, whereas 89.1% of the same-age 2001 children apprehended 3 directly; in addition, even the younger 2001 children rarely counted to name the number 3.

Table 8

For each number tested, percentage of 2001 children who used each naming strategy (C for counting and DA for direct apprehension), by age group (AG)

	AG					
Number tested	Strategy	3;3	3;9	4;3	4;9	5;3
2	C	2.2	6.3	2.7	1.3	1.25
	DA	97.8	93.7	97.3	98.7	98.75
3	C	16.7*	12.0	10.9	4.3	6.7
	DA	83.3*	88.0	89.1	95.7	93.3
4	C	100.0*	85.7*	66.7	38.8	35.0
	DA	0*	14.3*	33.3	61.2	65.0
5	C	/	100*	85.7	67.6	60.4
	DA	/	0*	14.3	32.4	39.6

Note. * The percentage pertains to less than 10 children.

Summary of results

The following main results stand out from the 1921-2001 comparison.

1 The difficulty order of the tasks was the same in 2001 as in 1921, the children used the same numerical evaluation strategies (all strategies described by Beckmann were found again and no new strategy emerged), and strategy changes with age (in the Naming task, transition from a counting strategy to a direct apprehension strategy) were the same today as they were in Beckmann's time.

2 On the other hand, the 2001 children outperformed the 1921 children on all four numerical tasks. This rise in scores corresponds to an advance in numerical development ranging from six months to one year, depending on the assessment test.

3 The magnitude of the performance gain varied across tasks. The greatest progress was observed on the Naming task (d=0.64), the smallest, on the Distinction task (d=0.19).

4 The superiority of the 2001 children was observed by the age of 3 years (on overall performance, on naming, and on recognition).

Discussion

In comparing cohorts separated by a period of 80 years, the risk of methodological bias is high and must be examined. Three questions will be discussed below: What are the consequences of a potential lack of representativeness of the 1921 sample? Were the 1921 children ill at ease or less familiar with the test situation than those of 2001? Do the tasks have the same meaning today as in Beckmann's time?

1 Although the 2001 sample was representative, the 1921 sample may not have been. However, only two possible cases exist: either the 1921 sample was skewed upwards, or it was skewed downwards. In the former case, the bias goes in the opposite direction to the hypothesis tested, which means that the observed rise would have been even greater if the 1921 sample had been representative. In the latter case, the bias goes in the direction of the hypothesis tested. Yet we noted in our study that the rise was still observed when the 2001 sample was confined to children from underprivileged homes. This allows us to contend that the possible lack of representativeness of Beckmann's sample does not invalidate the main results of our study.

2 Can the observed rise be explained in terms of the greater familiarity of today's children with testing situations, or by the fact that modern children feel more comfortable with adults? The young age of the children observed enables us to reject the first hypothesis. As for the second, nothing allows us to assume that the children observed in 1921, who were attending institutions where they were in contact with adults on a daily basis, were inhibited. Moreover, Beckmann eliminated from his sample any children who refused to be questioned. Besides, the potential consequences of inhibition in the 1921 children were limited, since the most obvious effect of social inhibition is a lengthening of response time, whereas all the tests were run without a time limit.

3 The remaining question concerns the meaning of the tasks. Do the tests used measure the same numerical abilities in 2001 as in 1921? Three signs allow us contend that they do: (1) the difficulty order of the four tasks was the same for the two cohorts, (2) the development curves in 1921 and in 2001 are similar (see Figures 2a and 2b), and (3) the two cohorts used the same types of strategies to evaluate numerical quantities. We can therefore legitimately rule out the idea that the rise in performance is a methodological artifact.

This rise was expected, insofar as the Flynn effect is a general phenomenon that has been observed in a variety of cognitive tasks (the Flynn effect was found again recently on Piaget's formal reasoning tasks; see Flieller, 1999). In addition, young children's scores on the numerical tests are correlated with their general cognitive development[5]. For example, the correlation between the numerical test scores (comparison of small quantities, counting, subtraction of 1, 2, or 3, etc.) and IQ on the WPPSI (Wechsler, 1972) averaged 0.52 for the 4;0 to 5;6 age groups. The observed rise on the four numerical tasks studied here can thus be regarded as a specific manifestation of the Flynn effect.

However, two unexpected findings in the present study can help us interpret and explain this phenomenon. The first pertains to the magnitude of the Flynn effect measured by numerical evaluation tasks. The rise in the total score on the four numerical tests over a period of 80 years corresponds to 6.4 IQ points (obtained by multiplying d by 15, i.e., the standard deviation of the standard scores on IQ scales). A 6-point increase over such a long period does not pose the interpretation problems encountered by Flynn (1998), who, based on successive standardizations of the Stanford-Binet and Wechsler tests, estimated a rise of 25 IQ points in White Americans between 1918 and 1995. The rise reported by this author is so great that it casts doubt on the validity of cross-cohort comparisons (e.g., a 1918 child with an IQ of 100 – the mean at that time – would have an IQ of only 75 points according to the 1995 norms). By contrast, an increase of 6 points in 80 years seems quite reasonable, given the many environmental and child-raising changes that occurred in the course of that period (better infant nutrition, more educated parents, smaller families on average, etc.; for a more complete analysis of these changes, see Flieller, 2001; Neisser, 1998; Storfer, 1990). If we compare this gain to the progress noted in French research on school children, another interesting fact appears. In a study on 7-year-olds tested using the French ECNI (*Echelle Collective de Niveau Intellectuel* or Collective Scale of Intellectual Level) over the 1973-1992 period, Flieller, Manciaux, and Kop (1994) found a rise of 1 point per decade for the verbal part and 3.8 points for the nonverbal part of the tests. In another study on 11-year-olds tested using the WISC-R (1981 norms) and WISC-III (1996 norms), the observed rise was 1.7 points per decade for the verbal part and 3.1 points for the nonverbal part (Wechsler, 1996). The increase we observed here in the numerical test scores (.8 points per decade) is thus closer to the verbal test increase than to that of the nonverbal tests. However, from a formal standpoint, our numerical tests are more like nonverbal tests, which are not characterized by verbal instructions but by responses expressed nonverbally. The only exception is the Naming task, which is entirely verbal, but this is precisely where the observed rise was the greatest. It thus seems that the form (verbal vs. nonverbal) of a test is not what is decisive, but rather the type of cognitive ability measured (fluid abilities for the nonverbal ECNI and WISC tests, crystallized abilities for the verbal and numerical tests; regarding this distinction, see Carroll, 1993). As far as the Flynn

effect is concerned, numerical tests behave like tests of crystallized abilities, i.e., skills acquired through interaction with the environment. If the environment changes, it is normal that these abilities will change too. This is exactly what a study by Genovese (2002) suggests. This author analyzed the content of tests taken by students entering the ninth grade in the State of Ohio between 1902 and 1913 and between 1997 and 1999. He found that the early 20th century tests called more upon factual knowledge, whereas the late 1990's tests necessitated greater mastery of relationships between concepts. According to Genovese, it is the change in what cognitive abilities are prized by society that accounts for the Flynn effect[6].

The second interesting and unexpected result obtained here is the age at which the rise was noted. The three-year-olds tested in 2001 obtained significantly higher scores than Beckmann's same-age children in 1921 (on the Recognition and Naming tests, and on overall performance). This difference was observed at too young an age to be explained by preschool education. Unexpectedly, the inter-cohort difference did not increase significantly with age, as could be expected if preschool had contributed to the higher scores. But school is not the only place where numerical knowledge is acquired. It is also acquired through cross-generational transmission, as Durkin, Shire, Riem, Crowther, and Rutter (1986) showed in their longitudinal study of verbal exchanges about numbers between mothers and their children under three. The explanation must therefore be sought in changes in child-raising practices. Although we do not have data here attesting to these changes, Flieller's (1996) work suggests that parental child-raising practices have indeed changed within the past few decades, that the changes are favorable to the cognitive development of children (seven-year-olds in that study), and that they can be partially accounted for by the fact that today's parents are more educated.

The observed rise in numerical abilities between 1921 and 2001 was not uniform but varied across tasks. We are unable to propose a satisfactory explanation for these variations. The best we can do is note that the smallest progress was observed on the Distinction test, which is the least likely skill to benefit from direct learning. While it is normal for a child to learn to count and name numbers, it is much less so for learning to distinguish between two numbers. This idea is supported by looking at the available preschool books (we did not find a single book containing exercises resembling the Distinction test). The greatest progress was noted on the Naming test, which is both the most difficult and the most verbal of the four tasks. If the above explanation of the overall rise is indeed correct, one can see why the positive effects of parent-child verbal interactions would show up the most on an entirely verbal test. We lack data for going further into this explanation. As Bideaud and Villette (1995) stressed, there are no systematic studies on child-raising practices in the area of numerical development, and it would be useful to "draw up an inventory of parental practices related to number learning and to analyze their impact on counting activities at various ages" (pp. 223-255; our translation).

What has this inter-cohort comparison taught us about numerical development? It stands out clearly from the results of this research that numerical development takes place at different paces, depending on the period under study (today's four- and five-year-olds are six to twelve months ahead of the children tested in 1921). Hence, numerical development is contingent upon the environment in which a child grows up, especially – if our interpretation is correct – upon the home environment. However, looking at the Naming task results, it seems that the mechanisms responsible for numerical development are relatively independent of educational practices in the home. Indeed, the strategies used to evaluate numerical quantities form a strict order (counting precedes direct apprehension via subitizing), and this order is maintained remarkably well over time. The impact of the home environment may therefore be that it can hasten the use of more mature strategies.

Conclusion

The present research makes an original contribution to understanding the Flynn effect. To our knowledge, no direct cohort comparisons have ever been made over such a long period. In addition, this is one of the rare studies that deals specifically with numerical development. It

revealed that the Flynn effect applies to number-evaluation tasks as well as to other types of tasks. Everything points to the conclusion that the rise in performance observed here corresponds to a genuine increase in the numerical knowledge of children between the ages of three and five and a half. However, the magnitude of the rise is much smaller than that observed by numerous authors for other tasks, and despite the lengthy period of time between the two cohorts tested, the rise remains within an acceptable range likely to result from the many changes occurring in Western societies. It seems, then, that the interpretation difficulties so often mentioned by Flynn – for whom performance gains on tests that have no observable equivalents in real life are not true gains – only pertain to tests that assess fluid abilities. Crystallized ability tests, on the contrary, bring out more moderate gains that seem to reflect real progress in cognitive abilities and to correspond in children to an acceleration of cognitive development. Preschool and school education is one of the factors often brought to bear (e.g., Emanuelsson & Svensson, 1990; Rowe, 1997; Teasdale & Owen, 2000) in accounting for higher cognitive test scores. This factor seems to be of little importance *here* since, as we have seen, the superiority of today's children was observed by the age of three and hardly increased in magnitude after that. The greater precociousness of today's children thus appears to be more the outcome of changes in child-raising practices. Although other authors have also forwarded this hypothesis (e.g., Emanuelsson & Svensson, 1990; Greenfield, 1998; Williams, 1998), further research is needed to prove that such changes have actually occurred. This interpretation is also in line with the role apparently played by cross-generation transmission in the initial acquisition of numerical knowledge. Whatever the case may be, the numerical development of children is moving faster today than it was eighty years ago, and traditional norms regarding the acquisition of numbers should be revised. The historical context clearly seems to have an impact on numerical development, but the processes at play in that development do not appear to have changed during the period considered here.

Notes

1 The measure of the increase in IQ points does not imply that the test is scored in IQ points: The difference of the means between two cohorts on any test can be divided by the standard deviation of the scores (Cohen's *d*) and then multiplied by 15 (standard deviation of the IQ's).

2 The choice of a regression without a constant term was justified by the fact that none of the children with a score of zero on the Production, Distinction, and Naming tests obtained a score different from zero on the Recognition test. If we had used a regression equation with a constant term, we would have artificially attributed these children with a non-zero recognition score.

3 The weighted group size was 42, which corresponds to the mean of the number of children in the five age groups.

4 Note that this order was maintained even when a child used only two of the strategies, and it was not refuted when a child used only one strategy.

5 For the 38 five-year-olds in his sample, Beckmann (1923) obtained a rank correlation of .93 between arithmetic scores (addition and subtraction) and the Binet-Simon mental age, but he did not calculate this correlation for the tasks considered here.

6 Genovese thinks that the emphasis placed on relations explains the rise in fluid abilities and the relative stagnation of crystallized abilities. But in our minds, his explanation of the higher cognitive performance brought about by cultural changes can be applied to both crystallized and fluid abilities.

References

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Beckmann, H. (1923). Die entwicklung der zahlleistung bei 2-6 jährigen kindern. *Zeitschrift für Angewandte Psychologie, 22*, 1-72 (French translation by J.-P. Fischer: *Le développement de la performance numérique chez des enfants de 2 à 6 ans.* Nancy: University Nancy 2).

Bideaud, J. (1999). La construction du nombre dans l'histoire des cultures humaines et chez l'enfant: Convergences et divergences. In G. Netchine-Grynberg (Ed.), *Développement et fonctionnement cognitifs: Vers une intégration* (pp. 197-218). Paris: Presses Universitaires de France.

Bideaud, J., & Villette, B. (1995). Les apprentissages numériques élémentaires: Psychogénèse polymorphe et variabilité interindividuelle. In J. Lautrey (Ed.), *Universel et différentiel en psychologie* (pp. 223-255). Paris: Presses Universitaires de France.

Bruner, J. (1996). *The culture of education*. Cambridge, MA: Harvard University Press.

Carroll, J.B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Descoeudres, A. (1921). *Le développement de l'enfant de deux à sept ans*. Neuchâtel: Delachaux & Niestlé (3rd edition, revised, 1946).

Durkin, K., Shire, B., Riem, R., Crowther, R.D., & Rutter, D.R. (1986). The social and linguistic context of early number word use. *British Journal of Developmental Psychology, 4*, 269-288.

Emanuelsson, I., & Svensson, A. (1990). Changes in intelligence over a quarter of a century. *Scandinavian Journal of Educational Research, 34*, 171-187.

Fischer, J.-P. (1991). Le subitizing et la discontinuité après 3. In J. Bideaud, C. Meljac, & J.-P. Fischer (Eds.), *Les chemins du nombre* (pp. 235-258). Lille: Presses Universitaires de Lille.

Flieller, A. (1996). Trends in child-raising practices: A partial explanation for the increase in children's scores on intelligence and cognitive development tests. *Polish Quarterly of Developmental Psychology, 2*, 51-61.

Flieller, A. (1999). Comparison of the development of formal thought in adolescent cohorts aged 10 to 15 years (1967-1996 and 1972-1993). *Developmental Psychology, 35*, 1048-1058.

Flieller, A. (2000). L'utilisation de données historiques en psychologie du développement. *Bulletin de Psychologie, 53*, 529-535.

Flieller, A. (2001). Problèmes et stratégies dans l'explication de l'effet Flynn. In M. Huteau (Ed.), *Les figures de l'intelligence* (pp. 43-66). Paris: Editions et Applications Psychologiques.

Flieller, A., Manciaux, M., & Kop, J.-L. (1994). Evolution des compétences cognitives des élèves en début de scolarité élémentaire sur une période de 20 ans. *Les dossiers d'éducation et formations, 47*, 205-218.

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

Flynn, J. R. (1994). IQ gains over time. In R.J. Sternberg (Ed.). *The encyclopedia of human intelligence* (pp. 617-623). New York: Macmillan.

Flynn, J.R. (1998). IQ gains over time: Toward finding the causes. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 25-66). Washington, DC: American Psychological Association.

Gelman, R. (1972). The nature and development of early number concepts. In H. Reese (Ed.), *Advances in child development and behaviour* (vol. 7, pp. 115-167). New York: Academic Press.

Gelman, R., & Gallistel, C.R. (1978). *The child's understanding of number*. Cambridge: Harvard University Press.

Genovese, J.E. (2002). Cognitive skills valued by educators: Historical content analysis of testing in Ohio. *The Journal of Educational Research, 96*, 101-114.

Greenfield, P.M. (1998). The cultural evolution of IQ. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 81-123). Washington, DC: American Psychological Association.

Kaufman, A.S., & Kaufman, N.L. (1993). *KABC: Batterie pour l'examen psychologique de l'enfant, Manuel d'administration et de cotation*. Paris: Editions du Centre de Psychologie Appliquée.

Neisser, U. (1997). Rising scores on intelligence tests. *American Scientist, 85*, 440-447.

Neisser, U. (Ed.). (1998). *The rising curve: Long-term gains in IQ and related measures*. Washington, DC: American Psychological Association.

Neisser, U. (1998). Introduction: Rising test scores and what they mean. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 3-22). Washington, DC: American Psychological Association.

Rodgers, J.L. (1998). A critique of the Flynn effect: Massive IQ gains, methodological artifacts, or both? *Intelligence, 26*, 337-348.

Rowe, D.C. (1997). A place at the policy table? Behavior genetics and estimate of family environmental effects on IQ. *Intelligence, 24*, 133-158.

Stern, W. (1927). *Psychologie der frühen kindheit bis zum sechsten lebensjahre* (4th ed. revised and enhanced). Leipzig: Quelle & Meyer.

Storfer, M.D. (1990). *Intelligence and giftness*. San Francisco: Jossey-Bass.

Teasdale, T.W., & Owen, D.R. (2000). Forty-year secular trends in cognitive abilities. *Intelligence, 28*, 115-120.

Wechsler, D. (1972). *Echelle d'intelligence de Wechsler pour la période préscolaire et primaire. W.P.P.S.I.* Paris: Editions du Centre de Psychologie Appliquée.

Wechsler, D. (1996). *WISC-III: Echelle d'intelligence de Wechsler pour enfants* (3ème ed.). Paris: Editions du Centre de Psychologie Appliquée.

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.

Cette recherche est une contribution au débat sur l'effet Flynn, consistant en une comparaison à très long terme (1921-2001) des connaissances numériques de deux cohortes d'enfants âgés de trois à cinq ans et demi. Beckmann (1923) a évalué le développement numérique des enfants à partir de quatre épreuves différentes (production, distinction, reconnaissance, dénomination). En 2001, ces mêmes épreuves ont été proposées à 400 enfants également répartis entre cinq groupes d'âge espacés de six mois. Les principaux résultats sont les suivants: (1) l'ordre de difficulté des quatre tâches est le même en 2001 qu'en 1921; (2) les performances de la cohorte de 2001 sont significativement supérieures à celles de la cohorte de 1921 (avance du développement numérique allant de six mois à un an selon les tâches); (3) la supériorité des enfants de 2001 s'observe dès l'âge de 3 ans; (4) la progression de scores varie selon les épreuves (la plus forte progression s'observe dans la tâche de dénomination, la plus verbale des quatre); (5) les enfants des deux cohortes utilisent les mêmes stratégies d'évaluation numérique mais, à âge égal, ceux de la cohorte de 2001 utilisent des stratégies plus matures dans la tâche de dénomination, en particulier le subitizing pour l'appréhension des petits nombres. La progression des scores semble correspondre à une réelle progression des compétences numériques, apparemment favorisée par les pratiques éducatives familiales (transmission intergénérationnelle).

Key words: Cohort comparison, Flynn effect, Intelligence, Numerical development, Subitizing.

Received: December 2002

Revision received: April 2004

Christine Bocéréan. GRAPCO, Laboratoire de Psychologie, Université Nancy 2, BP 3397, F-54015 Nancy Cedex, France. E-mail: Christine.Bocerean@univ-nancy2.fr

Current theme of research:

Analogical reasoning. Numerical development. Psychometrics. Semantic memory development.

Most relevant publications in the field of Psychology of Education:

Fischer J.P., & Bocéréan C. (à paraître). Impact de la réforme de 1970 sur les connaissances numériques des jeunes enfants. *Annales de Didactique et de Sciences Cognitives.*

Fischer, J.-P., & Bocéréan, C. (à paraître). Les modèles du développement numérique à l'épreuve de l'observation. *Bulletin de Psychologie.*

Jean-Paul Fischer. IUFM de Lorraine, 16 rue de la Victoire 57950 Montigny-les-Metz, France. E-mail: jfischer@lorraine.iufm.fr

Current theme of research:

Number development and numerical learning in children. Procedural knowledge vs. declarative knowledge.

Most relevant publications in the field of Psychology of Education:

Fischer J.P. (1992). *Apprentissages numériques: La distinction procédural/déclaratif.* Nancy: Presses Universitaires.

Fischer J.P. (1992). Subitizing: The discontinuity after three. In J. Bideaud, C. Meljac, & J.P. Fischer (Eds.), *Pathways to number* (pp.191-208). Hillsdale, NJ: Lawrence Erlbaum.

Fischer J.P. (1997). Tests implicites versus explicites et performances scolaires. *Psychologie & Education,* 31, 11-28.

Fischer J.P. (1998). La distinction procédural/déclaratif: Une application à l'étude de l'impact d'un "passage du cinq" au CP. *Revue Française de Pédagogie,* 122, 99-111.

Fischer J.P., & Bocéréan C. (à paraître). Impact de la réforme de 1970 sur les connaissances numériques des jeunes enfants. *Annales de Didactique et de Sciences Cognitives.*

André Flieller. GRAPCO, Laboratoire de Psychologie, Université Nancy 2, BP 3397, F-54015 Nancy Cedex, France. E-mail: Andre.Flieller@univ-nancy2.fr

Current theme of research:

Flynn effect. Intelligence. Psychometrics. Socio-cognitive interactions. Psychology of education.

Most relevant publications in the field of Psychology of Education:

Dickes, P., & Flieller, A. (Eds.). (2004, in press). Mesure et éducation. *Psychologie et Psychométrie* (special issue).

Flieller, A. (1999). Comparison of the development of formal thought in adolescent cohorts aged 10 to 15 years (1967-1996 and 1972-1993). *Developmental Psychology, 35*, 1048-1058.

Flieller, A. (1999). Les compétences et les performances cognitives dans l'évaluation scolaire. In J. Bourdon & C. Thélot (Eds.), *Education et formation: L'apport de la recherche aux politiques éducatives* (pp. 187-200). Paris: C.N.R.S. Editions.

Flieller, A. (2001). Problèmes et stratégies dans l'explication de l'effet Flynn. In M. Huteau (Ed.), *Les figures de l'intelligence* (pp. 43-66). Paris: Editions et Applications Psychologiques.