
André Flieller
University Nancy 2

Five standard Piagetian tests were administered to 180 adolescents between the ages of 10 and 15 years. The results were compared with those obtained in 1967 and in 1972 for similar participant samples. At equal ages, today’s adolescents exhibited a higher level of cognitive development than the adolescents of 20 or 30 years ago. The amount of gain observed varied across tasks, being very large for combinatory thought but mixed for conservation. This acceleration of cognitive development can partially explain the continuous rise in intelligence test performance (Flynn effect).

Many cohort comparisons have shown that the mean score on a variety of intelligence tests (e.g., Raven Progressive Matrices, Stanford–Binet Intelligence Scale, Wechsler Intelligence scales) has been rising by approximately 3 IQ points per decade for at least 60 years (see reviews by Flynn, 1984, 1987; Storfer, 1990). This rise has been noted in over 20 countries, including France, where the studies reported in the present article took place (e.g., Flieller, Manciaux, & Kop, 1995; Pry & Manderscheid, 1993; Wechsler, 1996).

Whereas the phenomenon of rising IQ scores, known as the Flynn effect, is well documented, it remains puzzling to many scholars, who are divided as to its interpretation (see Neisser, 1997). Many environmental variables have been proposed to explain the rise, including better nutrition (e.g., Lynn, 1990), the development of education (e.g., Teasdale & Owen, 1989), and the ever-increasing complexity of modern environments (e.g., Schooler, 1998). But none of these variables alone seems to be able to account for all of the observations, nor can any one variable explain performance changes of the magnitude observed. According to Flynn (1987, 1996, 1998), the observed gains are too great to correspond to a genuine increase in cognitive abilities. One of Flynn’s main arguments is that these massive IQ gains are not accompanied by similar progress in real life. A case in point is the fact that no improvement in scholastic achievement has been observed in the United States.

Some approaches aimed at better explaining the Flynn effect have recently been proposed. For Neisser (1997), the one-dimensional conception of intelligence that underlies the debate about the Flynn effect should be discarded: “Different forms of intelligence are developed by different kinds of experience . . . We are indeed very much smarter than our grandparents where visual analysis is concerned, but not with respect to other aspects of intelligence” (p. 447). On her side, Greenfield (1998) proposed attempting to determine the historical and cultural changes that might account for higher test scores and testing their explanatory power by means of field and lab experiments. To illustrate, she reported studies demonstrating the impact of video games and computers on the development of visual–spatial skills, measured on tests where a strong Flynn effect was found (e.g., the Object Assembly subtest of the Wechsler Intelligence Scale for Children [WISC]). According to Williams (1998), the Flynn effect is the result of multiple variables, whose effects should be studied separately. She advocated making finer-grained comparisons aimed at pinpointing specific changes that could be rooted in numerous variables.

These proposals all point to the utility of conducting analytic studies of performance improvements on cognitive tests. The above authors suggest that test-dependent or even item-dependent variations in the extent of the rise can provide researchers with cues about what variables are at play in the Flynn effect. The present article takes such an analytic approach. It reports two studies comparing the performance on Piagetian tasks of same-age adolescents from different cohorts. As shown below, Piagetian tasks have a number of original characteristics that are not found in conventional tests. As such, any specific performance changes observed on this type of task may help explain the Flynn effect. Note that cohorts have never been compared using Piagetian tasks, although certain observations suggest that the transition from the preoperational stage to the concrete operational stage is now occurring at a faster pace (Bronkhart as cited in Reese & McCluskey, 1984; Ruffy, 1981). This article focuses on the transition to formal operational intelligence. The choice of this developmental period was motivated by the importance of environmental variables in the acquisition of formal thought, a fact that Piaget (1972) himself acknowledged.

What are the characteristics of Piagetian tasks? It is now well established that the measures supplied by these tasks are correlated (sometimes strongly) with those obtained on conventional intelligence tests like the WISC or Raven Progressive Matrices. The correlation has been observed on formal-level tasks (Dupont, Gendre, & Pauli, 1975; Keating, 1975; Lim, 1988, 1996; Longeot, 1969; Neimark, 1975) as well as on concrete-level ones. Cloutier

I would like to thank Francois Longeot and Jacqueline Hornemann for the information they so willingly provided; Delphine Lambert and Yousef Tazouti for their valuable collaboration; Charles Brainerd and Douglas Detterman for their comments on an earlier version of this article; and Vivian Waltz for the English translation.

Correspondence concerning this article should be addressed to André Flieller, Laboratory of Psychology, University Nancy 2, B.P. 33-97, F-54015 Nancy Cedex, France. Electronic mail may be sent to flieller@clsh.univ-nancy2.fr.
and Goldschmid (1976), for example, found a correlation of .46 between a formal operations test (proportion test) and the Raven Progressive Matrices. The exact factorial composition of Piaget’s tasks is not known because of a number of methodological problems, particularly concerning task sampling (see reviews by Carroll, 1993; Grégoire, 1991; Lautrey, Ribaupierre, & Rieben, 1990). Some conclusions may nevertheless be drawn. A general and often strong factor has been found in every factor analysis of Piagetian tasks (e.g., Jensen, 1980). However, group factors are also noted. For example, several studies (Inman & Secrest, 1981; Lautrey et al., 1990; Londeix, 1985) found evidence of an opposition between spatial reasoning tasks and logical reasoning tasks (classification, seriation, quantification of probabilities, propositional logic, etc.). Likewise, a factor specific to conservation tasks has been found in some studies (Carroll, Kohlberg, & De Vries, 1984; Lautrey et al., 1990). The relations between these factors and those of conventional tests need to be clarified (Carroll, 1993). One can say, however, that spatial tasks, Piagetian or otherwise, have their own specificities. Moreover, a few studies (e.g., Humphreys, Rich, & Davey, 1985) have shown that Piagetian tasks have a higher loading on the nonverbal factor than on the verbal factor, and certain authors (e.g., Rubin, Brown, & Pridde, 1978) take Piagetian measures to be akin to measures of fluid intelligence.

Despite these similarities, Piagetian tasks have certain interesting properties. First, they do not have time limits; performance thus depends on reasoning ability, not execution speed. This point is important, because a number of authors (e.g., Brand 1987a, 1987b) have explained the Flynn effect in terms of an increase in responding speed, and Flynn himself (1996), although he rejected this hypothesis, felt that it was “theoretically ideal” (p. 24). Second, a correct response is only accepted in Piagetian tasks if the participant is capable of justifying it. Although the validity of using children’s explanations as a criterion for assessing their responses has been debated (see Brainerd, 1973a, 1977b), children’s justifications can provide useful information in diachronic comparisons. In particular, they may reveal changes in problem-solving strategies. Third, there is a hierarchical order among the items in a given task that has been validated using various techniques (hierarchical analysis, Rasch analysis; see Bond, 1995; Green, Ford, & Flamer, 1971). The existence of this hierarchy brings out potential changes in the order of the items, and any change of this type can be taken as an indication of a potential variation over time in the meaning of the items concerned. Last, Piagetian tasks measure the acquisition of specific concepts and ways of reasoning, such as the concept of proportion at the formal operational level. As a consequence, accurate detection of cognitive progress contributing to the Flynn effect is likely to be easier than with traditional tests.

It should be noted that the above description does not imply adherence to Piaget’s theory. Moreover, except perhaps for the second point, Piagetian tasks are not the only ones to have the features mentioned above (e.g., like Piagetian items, many items in the Stanford-Binet test are not timed). However, only a few tests possess all four properties, and they have rarely been used to compare cohorts. Piagetian tasks thus have something original to offer, and this puts them in a good position for supplying new data to account for the Flynn effect.

What can one expect to find if one compares the responses of adolescents on the formal operations tasks to those of same-age adolescents 20 years earlier? What might be the significance and utility of such observations?

If the rise in IQ scores corresponds in part to a faster pace of cognitive development, this should come through on the Piagetian tasks. In other words, access to the formal operations level should occur earlier today than 20 years ago. Improvements in performance on Piagetian tasks cannot be explained in terms of faster processing speed or more intelligent guessing, two variables suggested by Brand (1987a, 1987b). Nor can an overall increase in formal operations task scores be the result of the development of specific visual analysis skills (Neisser, 1997), because some of these tasks do not tap this type of ability (e.g., quantification of probabilities).

However, the Piagetian tasks are not homogeneous, as some factor analytic studies have shown, and they are affected to different extents by past experience (for the formal operational level, see Larivée, Longeot, & Normandeau, 1989). An example is performance on conservation tasks, which seems to be less affected by the level of schooling than other tasks, such as seriation or quantification of probabilities (Laurendeau-Bendavid, 1977). By contrast, combinational thinking problems, such as the permutation of tokens, seem to be highly dependent on the mathematics curricula offered in the schools (Longeot, 1968). If performance has truly improved more on tests with a visuospatial component than it has on other tests (Flieller, Jautz, & Kop, 1989; Lynn & Hampson, 1986), then this effect should show up on some Piagetian tasks. For instance, the so-called mechanical curves problem (Nassef, 1963), which relies on spatial representations, should give rise to greater cross-cohort differences than the other formal operations tasks.

These trends in access to formal thinking are not relevant to the Flynn effect alone but also contribute to the debate on the learning of Piagetian concepts, a debate that reached its peak in the 1970s (reviews reflecting different points of view have been published by several authors, e.g., Brainerd, 1973b, 1977a; Kuhn, 1974; Lefebvre-Pinard, 1976; Strauss, 1972, 1974). Some Piagetian authors contend that improvements in performance obtained in a laboratory setting cannot be equated with the spontaneous progress observed during development. One of the merits of cohort comparisons is that they bring out differences obtained under natural conditions. Of course it remains to be shown that these differences are not completely artifactual, which brings one back to the controversy about the Flynn effect.

If rising test scores reflect faster cognitive development, the responses should be qualitatively the same for a given performance level: Correct answers should be justified by the same arguments and even result from the same strategies, and incorrect answers should be caused by the same reasoning errors. Qualitative differences in the responses may be indicative of changes in the procedures used, making cohort comparisons problematic. Suppose, for example, that only the adolescents in the most recent cohort use a mathematical formula to solve a combination problem. Such a difference could hardly be regarded as a developmental difference.

It also follows from the hypothesized speed-up in the pace of cognitive development that the item hierarchy for a given task will not vary. The fact that, when ranked by difficulty, the order of the items has not changed over time obviously does not suffice to prove the hypothesis. On the other hand, a significant modification
in this order could lead one to suspect changes in the very meaning of the measures.

Surveys on the level of operational reasoning in adolescents that could potentially serve as a basis for cohort comparisons are scarce. Such studies would have to meet a number of conditions regarding the samples and tasks, and these conditions are rarely satisfied. In France, the only adequate data can be found in two norming studies on the Logical Thought Development Scale. This scale was first standardized in 1967 by its author, who published norms for each task (Longeot, 1967). Because the norming groups used by Longeot were small (30 participants per age group), Hornemann standardized Longeot’s scale in 1972 on a larger sample; however, this sample was limited to 10- to 12-year-olds, and norms were constructed only for the total score.

Therefore, I conducted two comparative studies. The first study was based on the 1972 sample. Its specific aims were to test the hypothesis of an improvement in overall performance in formal operational tasks and to compare this improvement with that observed on conventional tests. The second study was based on the 1967 sample. It was aimed at duplicating the previous cohort comparison on 13- to 15-year-olds by studying performance changes task by task and making precise, qualitative comparisons of the responses on certain items.

Given that both studies were based on Longeot’s scale and that this test is not well-known outside of France, I describe the scale in the first part of this article, which serves as a general introduction to the present investigations.

**Longeot’s Logical Thought Development Scale**

Longeot’s Echelle de Développement de la Pensée Logique (Logical Thought Development Scale [LTDS]) was designed to measure the cognitive development of adolescents (see Longeot, 1967, 1969). It comprises five tests, which are administered individually. Each test corresponds to the standardization of a problem used previously by Piaget and Inhelder in their experiments (e.g., Inhelder & Piaget, 1958; Piaget & Inhelder, 1951; Piaget, Inhelder, & Szeminska, 1956).

In compliance with Piagetian tradition, the operational level (concrete, or C; intermediate, or I; formal A, or Fa; formal B, or Fb) is assessed not only according to the children’s ability to solve the test problems but also on the basis of the explanations they give to support their answers and their reactions to the experimenter’s objections and countersuggestions. However, unlike Piaget’s clinical method, the experimenter’s interventions are strictly defined in the test manual.

**Conservation**

The materials consist of three same-size balls—two of clay and one of metal—and two containers of water. There are three items in the test. The first assesses the dissociation of weight and volume. The experimenter asks the participant to explain why the water rises when the balls of clay are dropped in the water (explanation by weight or by volume) and to predict the height of the water in the containers (equal or unequal levels). Next, the experimenter shows the participant one of the clay balls and the metal ball and asks him or her to predict the water levels in the containers and justify the answer given. The second item tests conservation of volume. The experimenter rolls a ball of clay into a long sausage shape and asks the participant to predict the water levels if the ball of clay is dropped in one container and the sausage-shaped clay is dropped in the other. A justification is requested. The same questions are asked after a second transformation in which the sausage is broken up into 10 or so pieces. Objections are made if the answer is incorrect. The third item tests conservation of weight. The participant must predict the relative weights of the untransformed ball of clay and the other ball of clay after it is flattened and then broken into small pieces. Justifications are requested, and objections are made to incorrect answers. The third item is given only following failure on both of the preceding items; otherwise, the participant is credited with a success.

**Level C** is defined as success on conservation of weight only, and **Level I** is defined as success on conservation of volume or on volume-weight dissociation. Participants who succeed on both conservation of volume and volume-weight dissociation are defined as intermediate or better (≥I).

**Permutation**

The materials are tokens of different colors. The test has four items. In the first, the participant has to predict how many different ways a yellow token, a red token, and a blue token can be lined up (prediction). The participant is then asked to line up the tokens on the table in all possible ways (execution). The experimenter makes suggestions in cases of failure. Only the execution phase is scored. Then a green token is added and the same procedure is followed (prediction followed by execution). The prediction of the number of permutations is the second item, and execution is the third. For the fourth item, still another token is added (purple) and the participant is simply asked to predict the number of possible permutations of the five colors. The participant must explain the method used, and the experimenter determines whether the participant is capable of generalizing the method to six and seven colors.

The scoring is based on response accuracy, the number of suggestions made by the experimenter, and the method used by the participant (e.g., the answer 12 on Item 2 is not accepted if it is obtained by doubling the permutations of three colors: 6 permutations for 3 colors, therefore 12 permutations for 4 colors).

On this test, Levels C and I are not distinguished: Participants who succeed on the first item are classified at the concrete or intermediate level (C–I), and those who fail on this item are classified as below C (<C). Level Fa is defined as success on Item 2 or 3, and Level Fb is defined by success on Item 4.

**Quantification of Probabilities**

In this test, two kinds of tokens are used: plain yellow tokens and yellow tokens with a large black X on one side. Each item uses two collections of tokens; for example, a collection of four tokens, one of which has an X, and a collection of four tokens, two of which have an X. The participant has to decide in which of the two collections he or she will have a greater chance of picking a token with an X on the first try, without looking.

The test includes eight items. In the four Level-C items, the participant has to compare the following proportions (where the numerator is the number of tokens with an X and the denominator is the total number of tokens): ¼ and ²⁄₅; ⁵⁄₈ and ³⁄₄; ³⁄₈ and ½; ¼
and \( \frac{1}{2} \). The level is reached if two or more of these problems are correctly solved. In the single Level-I item, the proportions \( \frac{1}{2} \) and \( \frac{1}{4} \) have to be compared. In the two Level-Fa items, \( \frac{3}{6} \) has to be compared with \( \frac{1}{3} \), and \( \frac{1}{6} \) with \( \frac{1}{4} \). The Fa level is reached if at least one problem is correctly solved. Finally, the one Level-Fb item involves the comparison of the proportions \( \frac{1}{6} \) and \( \frac{1}{3} \).

**Pendulum Oscillations**

The materials consist of a stand, a hanging string, and five weights (50 g, 100 g, 150 g, 200 g, 250 g) that can be hung on the string. The experimenter hangs a weight, sets it in motion, and then explains to the participant that different variables can affect the period of the pendulum: the length of the string, the weight of the pendulum, the force of the initial push, and the height from which the pendulum is released. The participant is then asked to do experiments to determine which variables account for the pendulum’s motion and is allowed to perform whatever experiments he or she wishes. If the participant changes only one variable at a time, the experimenter proposes changing two—for, example, the weight and the length of the string (countersuggestion). When the participant has finished, the experimenter asks him or her to draw conclusions.

This test, which has only one item, is a Level-Fb test. This level is reached if and only if the participant proceeds in a systematic way by changing only one variable at a time. In cases of failure, the participant is rated as below Fb (<Fb).

**Mechanical Curves**

This test requires a small apparatus composed of a cylinder fastened to a stand that can be rotated around its horizontal axis using a crank. The cylinder is covered with a standard sheet of paper (8 × 11\( \frac{1}{2} \) in.). A pencil can be moved horizontally from one end of the cylinder to the other along a bar attached on top of the stand. The task consists of imagining and drawing the line drawn by the pencil as it moves along the rotating cylinder. Thus, two simultaneous movements must be coordinated: the horizontal motion of the pencil and the rotation of the cylinder.

The test has six items: the pencil moves without cylinder rotation (Level C); the cylinder rotates without pencil movement (Level C); the cylinder rotates and the pencil moves from one end to the other (Level I); the cylinder rotates and the pencil goes and comes back (Level Fa); the cylinder does two full rotations and the pencil goes and comes back (Level Fa); and the cylinder does two rotations and the pencil goes to one end but does not come back (Level Fb).

Level C is reached when Item 1 or 2 is solved, Level I when Item 3 is solved, Level Fu when Item 4 or 5 is solved, and Level Fb when Item 6 is solved.

**Total Score and Overall Operational Level**

To determine the total score, one must weight the items in such a way that the different tests have the same weight at each operational level. For example, the conservation test and the permutation test have a single Level-C item, whereas the probability test has four and the mechanical curves test has two. Thus, success on these items is scored 2 on conservation and permutation, \( \frac{1}{2} \) on probability, and 1 on mechanical curves.

In the Piagetian perspective, the total score is of little direct interest. As a consequence, it is converted into an overall operational level by means of a table. The resulting operational levels are C, I, Fa, and Fb. In a psychometric perspective, the total score obviously has a value of its own, and it is particularly interesting in cohort comparisons. For additional information on the LTDS, readers may wish to refer directly to Longeot’s publications (1966, 1967, 1969).

**Study 1**

The first study was designed to compare the Longeot test scores of adolescents aged 10–13 years examined in 1972 versus 1993.

**Method**

**Participants**

1972 participants. Following the publication of the LTDS (Longeot, 1967), Hornemann established the test norms using a larger sample than Longeot’s. Data collection took several years and was completed in 1972 (J. Hornemann, personal communication, September 18, 1995). Her results were included in the 1974 LTDS manual. The sample was selected using the quota method and was representative of the French school population at the time. It consisted of three age groups. The 10-year-old group included 150 participants ranging in age from 9 years 6 months to 10 years 5 months; the 11-year-old group included 180 participants who ranged in age from 10 years 6 months to 11 years 5 months; and the 12-year-old group consisted of 180 participants ranging in age from 11 years 6 months to 12 years 5 months.

1993 participants. The 1993 sample was drawn from nine elementary schools and three junior high schools in the metropolitan area of Nancy, France. It included three age groups with 30 participants each: a 10-year-old group (ranging in age from 9 years 6 months to 10 years 5 months), an 11-year-old group (ranging in age from 10 years 6 months to 11 years 5 months), and a 12-year-old group (ranging from 11 years 6 months to 12 years 5 months). The age groups were defined as in Hornemann, but the medians were slightly below the theoretical ones (9 years 9 months, 10 years 11 months, and 11 years 10 months instead of 10 years, 11 years, and 12 years).

A weighting procedure using the weight command of SPSS (Statistical Package for the Social Sciences) was applied to each subsample (age group) in order to ensure representativeness with respect to national quotas. Four variables were controlled: sex, nationality, school grade, and father’s occupation. In other words, for each age group, the distribution of the students along these four variables was the same in the sample and in France as a whole.

**Scale and Scoring**

The study used the LTDS scale described above. In 1993, as well as in 1972, the tests were scored by an assistant and checked by me, and I adhered strictly to the scoring instructions in the manual. These instructions, which have not been changed since the publication of the LTDS, are very clear, and no interpretation problems were encountered.

**Procedure**

In 1972, the LTDS was administered in accordance with the standard procedure (full administration of the scale). In 1993, an abbreviated version had to be used in order to comply with the 1-hr-per-student time limit set.
by the schools, where the adolescents were tested individually. To save time, I did not give the pendulum test because 10- to 12-year-old children rarely succeed on this problem. For the same reason, the permutation test was stopped when the participant failed the \( Fa \)-level item, and the probability and mechanical curves tests were stopped when the participant failed on the \( l \)-level item. The procedure followed in 1993 on these three tests is authorized in the LTDS manual. In other words, the dropped items are considered optional in the manual, although they were in fact administered in 1972.

Results

Statistical Analyses

Because of the small sample size in 1993 (30 per age group), all statistical processing was done with and without weighting. The results with weighted and unweighted procedures yielded no substantive differences given that the quotas were controlled as strictly as possible during data collection. The results presented below were obtained with weighted statistical procedures. (The results of unweighted statistical procedures are available from the author.)

Psychometric Properties of the LTDS in 1993

Principal-components analysis of the LTDS. When standardizing the LTDS, Homemann did not study the psychometric properties of the scale. To fill this gap, I analyzed the structure of the LTDS using a principal-components analysis on the 1993 sample. On the basis of Kaiser’s criterion, only one factor could be retained, which accounted for 56% of the total variance. The factor loadings were .59 for conservation, .80 for permutation, .82 for probability quantification, and .77 for mechanical curves.

Reliability of the LTDS tests. The reliability of the LTDS tests was measured by Cronbach’s coefficient alpha. The following values were obtained: .62 for conservation (three items), .63 for permutation (four items), .81 for probability quantification (eight items), .66 for mechanical curves (six items), and .83 for the whole scale.

Total Score and Operational Levels in Each Cohort

The LTDS manual supplies two types of data taken from Homemann’s standardization of the test: the distribution of the total score in the 1972 sample and the frequency table of the overall operational level in that same sample.

Comparison of the total scores in 1972 and 1993 did not yield a significant difference in the variance: In 1972, \( SD = 4.6; \) in 1993, \( SD = 4.8; \) Levene’s test for equality of variance, \( F(1, 598) = 0.07, p = .80. \) In contrast, the mean of the total score increased significantly, rising from 11.3 in 1972 to 12.4 in 1993, \( t(598) = 2.04, p < .03 \) (one-tailed). The differences in the means were equivalent at 0.23 \( SD \). For the normalized scores \( (M = 100, SD = 15) \), the progression was thus 3.5 points over 21 years, or 5 points in a 30-year generation. No age-linked variation in this progression was found: Cohort \( \times \) Age, tested by an analysis of variance, \( F(2, 594) = 28.01, p = .26. \)

Table 1 gives the distribution of participants across the four operational levels by testing date. In 1993, the percentage of 10- to 12-year-olds at Level C was lower in 1993 than in 1972: after

<table>
<thead>
<tr>
<th>Level</th>
<th>Age 10–12 years (Study 1)</th>
<th>Age 13–15 years (Study 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1972 (( n = 510 ))</td>
<td>1993 (( n = 90 ))</td>
</tr>
<tr>
<td></td>
<td>1967 (( n = 90 ))</td>
<td>1996 (( n = 90 ))</td>
</tr>
<tr>
<td>Concrete</td>
<td>46</td>
<td>37</td>
</tr>
<tr>
<td>Intermediate</td>
<td>44</td>
<td>47</td>
</tr>
<tr>
<td>Formal A</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>Formal B</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
| Note. Percentages have been rounded off. LTDS = Echelle de Développement de la Pensee Logique (Logical Thought Development Scale).

Discussion

The use of an abbreviated version of the LTDS in 1993 slightly reduced cross-cohort differences, which would have been a little larger if the pendulum test had been given and if the other tests had been administered in their entirety. Therefore, the above results are conservative.

The improvement in performance found in the present study (5 points in a generation) is not as great as that observed in nonverbal tests from three other French studies on children (to facilitate comparisons, I standardized the scores and calculated the progression over a 30-year period). Pry and Manderscheid (1993) noted a 15.2-point progression in one generation for 9- to 10-year-old children tested on the Raven Coloured Progressive Matrices in 1962 and 1993. For 7-year-olds tested in 1973 and 1992, Flieller et al. (1995) observed a gain of 11.4 points on the nonverbal part of a time-limited group test, the Echelle Collective de Niveau Intellectuel (Collective Scale of Intellectual Level [ECNI]), and 3.1 points on the verbal part. A comparison of the scores obtained by the same sample of 11-year-old children on the revised version of the WISC (1981 norms) and on the 3rd edition of the WISC (WISC–III; 1996 norms) revealed a gain in one generation of 9.2 IQ points on the Performance scale and 5 IQ points on the Verbal scale (Wechsler, 1996). The difference between the LTDS results and the nonverbal test results in these three studies is all the more notable because, as I indicated earlier, Piagetian tasks are more strongly correlated with nonverbal tests than with verbal ones. The difference may be due to the fact that visuospatial skills play a lesser role in the LTDS than in the nonverbal tests in question. This interpretation is consistent with Neisser’s (1997) explanation of the Flynn effect. Regarding the execution-speed account proposed by Brand (1987a, 1987b), speed seems to play a secondary role in the performance gains because they were as great on the LTDS and on the WISC–III Verbal scale (both untimed) as on the ECNI verbal tests (timed).

Study 2

The second study was primarily aimed at extending the comparison to older adolescents (aged 13 to 15 years) and examining

Table 1

Overall Level of Development on the LTDS by Age and Date of Administration

<table>
<thead>
<tr>
<th>Level</th>
<th>1972 (( n = 510 ))</th>
<th>1993 (( n = 90 ))</th>
<th>1967 (( n = 90 ))</th>
<th>1996 (( n = 90 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concrete</td>
<td>46</td>
<td>37</td>
<td>22</td>
<td>12</td>
</tr>
<tr>
<td>Intermediate</td>
<td>44</td>
<td>47</td>
<td>43</td>
<td>34</td>
</tr>
<tr>
<td>Formal A</td>
<td>9</td>
<td>13</td>
<td>26</td>
<td>40</td>
</tr>
<tr>
<td>Formal B</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>15</td>
</tr>
</tbody>
</table>

Note. Percentages have been rounded off. LTDS = Echelle de Développement de la Pensee Logique (Logical Thought Development Scale).
the performance trends on each LTDS test, because this could not be done with the results published by Hornemann. The data used were collected by Longeot himself in his investigation of the LTDS.

Method

Participants

1967 participants. In Longeot’s (1967) statistical analysis of the LTDS, 210 participants between the ages of 9 and 15 years were examined (30 per age group). In the present study, only participants aged 13, 14, and 15 years were considered. The 10- to 12-year-old participants from Study 1 could not be included in Study 2, because the full LTDS was used in the second study whereas an abbreviated version of it was used in the first. The age ranges in the groups were defined as in Hornemann (specified by F. Longeot in a personal communication, December 18, 1993): The 13-year-old age group included participants aged 12 years 6 months to 13 years 5 months; the 14-year-old age group included those aged 13 years 6 months to 14 years 5 months; and the 15-year-old age group included those aged 14 years 6 months to 15 years 5 months (with a few participants aged 15 years 6 months to 16 years 5 months).

The tests were administered in schools located in the Paris area. Longeot had matched each age group to its national distribution by school grade, thus controlling an important variable. However, because of the specific geographical area in which the students lived, the representativeness of the sample cannot be fully guaranteed. Nonetheless, no significant difference was found between Longeot’s Parisian sample and Hornemann’s national sample in the operational level distributions of the 10- to 12-year-olds: After combining groups Fa and Fb, $\chi^2(2, N = 600) = 0.43, p > .80$. Moreover, if a bias did exist in the sample, it would certainly skew the data upward (making the mean score higher in the sample than in the population), because the Paris area is socially more privileged than the rest of France. Such a bias runs counter to the hypothesis tested here.

1996 participants. The participants in 1996 came from four junior high schools in the metropolitan area of Nancy, France. They were 13 ($n = 30$), 14 ($n = 30$), and 15 ($n = 30$) years old. The age ranges were the same as in Longeot (1967), but the medians were slightly below the theoretical ones (12 years 11 months, 13 years 8 months, and 14 years 11 months instead of 13 years, 14 years, and 15 years).

As with the 1993 sample, I applied a weighting procedure (using the SPSS weight command) to the three subsamples (age groups) in order to represent the national quotas for sex, nationality, school grade, and father’s occupation.

Scale and Scoring

In Study 2, the tests were scored by an assistant and checked by me, as in Study 1. This method is similar to the one used in 1967, where the tests were either scored by Longeot alone or by an assistant first and then again by Longeot. The instructions in the LTDS manual, written by Longeot, were strictly followed. It is worth noting that in a secondary analysis of Longeot’s data, Lautrey (1980b, pp. 686) scored the tests of the 1967 sample himself. He found a high degree of reliability (although he did not mention the reliability coefficient).

Procedure

The entire LTDS was administered in 1996. Thus, as in 1967, the participants took all five LTDS tests and all items in each test. The adolescents were tested in their schools both in the 1967 study and in the 1996 study.
tested in that year. For 1993, the use of a stopping criterion prevented calculation of the interstage indexes, where there may have been cases of participants who, say, failed on all Level-I items of a given test but would have succeeded on a Level-F item had they been tested for that level. The hierarchical analysis of the LTDS was thus applied to the 1996 sample alone.

Looking at Table 2, one can see that the interstage index values in 1967 and 1996 are very close except for conservation, where the index is at its maximum in 1996 (at 13–15 years, the probability of making an interstage error on this test was virtually zero). These very high values for the interstage index indicate that the LTDS tests conform to the stage model, both in 1967 and in 1996.

The value of the intrastage index was higher in 1996 than in 1967 on conservation, quantification of probabilities, and mechanical curves. This means that for these tests, items at the same operational level tended to develop in a fixed order in 1996 but not in 1967. In contrast, the intrastage index was lower in 1996 on the permutation test. In 1967, the correct prediction of the number of permutations of four tokens tended to be subordinate to the ability to produce the permutations; this was no longer true in 1996.

Operational Levels in Each Cohort

Table 1 gives the distribution of participants across the four overall operational levels. The percentage of 13- to 15-year-olds at Level Fa or Fb was significantly higher in 1996 (55%) than in 1967 (35%), \( \chi^2(3, N = 180) = 8.02, p < .03 \) (one-tailed).

Table 3 gives the participants’ operational levels on each test. On the conservation test, the operational levels of the participants were significantly different in 1967 and 1996, \( \chi^2(3, N = 180) = 7.56, p < .03 \) (two-tailed). There was both a progression (increase in the percentage of participants at Level C and increase in the percentage of participants at Level I) and a regression (drop in the percentage of participants at Level ≥I). In one of his publications, Longeot (1966) mentioned that the weight–volume dissociation item was correctly solved by 50% of the 13-year-old participants, which makes a success rate of at least 63% at age 13–15 years. This percentage was only 47 in 1996, \( \chi^2(1, N = 180) = 5.05, p < .05 \) (two-tailed). Because the volume conservation item seems to have been answered as well in 1996 as 30 years ago (Longeot found 70% success at age 12 years, and I found 84% success at age 13–15 years in 1996), it is the drop in performance on the weight–volume dissociation item that is responsible for the lower percentage of ≥I participants on the conservation test. Note that 38% of those who failed on this item had attained the formal level (Fa or Fb) on the LTDS as a whole.

Improvement was observed on the other four tests. The percentage of participants at the formal level (Fa or Fb) was higher in 1996 than in 1967 on permutation, \( \chi^2(3, N = 180) = 39.97, p < .00001 \) (two-tailed); quantification of probabilities, \( \chi^2(3, N = 180) = 11.17, p < .02 \) (two-tailed); the pendulum, \( \chi^2(3, N = 180) = 4.12, p < .05 \) (two-tailed); and mechanical curves, \( \chi^2(3, N = 180) = 9.74, p < .03 \) (two-tailed).

Cramer’s \( V \) coefficient can be used to compare intercohort differences across tests. \( V = .20 \) for conservation, .48 for permutation, .25 for probabilities, .15 for the pendulum, and .23 for mechanical curves.

Qualitative Study of Responses

Because the 1967 protocols were destroyed, a complete qualitative study of the responses was not possible. Some data for comparisons are nevertheless available.

The LTDS manual and Longeot’s publications contain a great deal of information about the responses obtained in 1967 on the various LTDS items, both for correct answers (justifications) and for incorrect answers (error types). This information was used by both Longeot and myself to categorize and score the responses. All response types observed in 1996 were also found by Longeot because no categorization problems arose in this study (this remark also applies to Study 1). On the weight–volume dissociation item,
for example, the adolescents who failed, whether in 1996 or in 1967, thought that the rise in the water level was a function of the object's weight, not of its volume. The only exception occurred for the fourth item of the permutation test (prediction of the permutations of five tokens of different colors), where 1 adolescent applied the mathematical formula \( P_n = n^2 \). However, this was a special case (both of this girl's parents were mathematics teachers). The other students who succeeded on this item discovered the solution using the types of reasoning observed by Longeot (e.g., taking the result previously found for four colors and multiplying it by five).

However, were the frequencies of the various types of responses the same as 30 years ago? A partial answer to this question can be found in the detailed information about two items supplied by Longeot (1967). For the first item (the prediction of the permutations of four tokens), Longeot found six types of answers. Two (the answer 8 and the answer 12) were generalizations of the solution to the preceding three-token permutation problem (\( 6 = 3 \times 2 \)): 8 was obtained by multiplying 4 colors by 2, and 12 was obtained by multiplying the result for three colors also by 2 (\( 12 = 6 \times 2 \)). These are Level-C responses, as are the other incorrect answers (e.g., "as many ways as colors," in this case, 4). The remaining three response types were Level Fa, which added 1 point to the score. The correct answer (24) was relatively infrequent at the ages considered here. The answers 12 and 16, although incorrect, reflected multiplicative reasoning at the formal level (\( 12 = 4 \times 3 \) and \( 16 = 4 \times 4 \)). Table 4 can be used to compare the response distributions on this item in 1967 and 1996. One can see that the distributions are roughly the same for the two cohorts: The answers 12 and 16, although incorrect, added 1 point to the score. The correct answer (24) was relatively infrequent at the ages considered here. The answers 12 and 16, although incorrect, reflected multiplicative reasoning at the formal level (\( 12 = 4 \times 3 \) and \( 16 = 4 \times 4 \)).

The other item studied by Longeot was the third mechanical curve problem, in which the participant was asked to draw the line connecting the two corners of the paper (4 colors, 2 movements). The answer 12 was clearly the predominant response. Participants who incorrectly drew two perpendicular lines simply juxtaposed the two movements. Participants who incorrectly drew a sinusoidal curve attempted to compound the two movements. The fourth category of incorrect responses included various lines that moved in the right direction (from the upper left-hand corner to the lower right-hand corner of the paper) but were not straight (wavy line, series of horizontal lines, etc.). The distributions of the response types in 1967 and 1996 are given in Table 5. One can see again that the two cohorts are close together on the incorrect answer categories. The only notable difference is the presence of a response not mentioned by Longeot in 1967—namely, a curved line connecting the upper-left and lower-right corners of the paper. But this response is merely a variant of the fourth category described above.

\[ \text{Table 4} \]

<table>
<thead>
<tr>
<th>Response</th>
<th>1967</th>
<th>1996</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level-C answers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>40</td>
<td>22</td>
</tr>
<tr>
<td>12 (6 \times 2)</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Other Level-C answers</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Level-Fa answers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 (4 \times 3)</td>
<td>21</td>
<td>36</td>
</tr>
<tr>
<td>16</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>24</td>
<td>8</td>
<td>14</td>
</tr>
</tbody>
</table>

**Discussion**

This study showed that the mean operational level of the 13- to 15-year-old adolescents was higher in 1996 than in 1967. This finding confirms the main result of the first study (note that the 10- to 12-year-old participants tested in 1993 were born in 1983, 1982, and 1981 as were the 13- to 15-year-old participants tested in 1996). However, compared with Study 1, this study has two limitations. First, it was not possible to compare the LTDS scores because Longeot only published the overall operational level of his participants. Second, the 1967 sample was not as representative as the national sample of 1972. Nevertheless, given that the distributions of the 10- to 12-year-olds' operational levels were the same in the two samples (as indicated above in the Method section), it seems legitimate to use the 1967 sample as a base for comparison.

The operational level of the 13- to 15-year-olds rose to variable extents on the different LTDS tests. The greatest increase occurred on the permutation test. Because of the small number of items (four), this test was not very reliable. Even so, the low reliability level cannot account for this particular trend because the changes observed between 1967 and 1996 are one-directional. Another possible reason for the rise could be related to school learning, although the mathematics curriculum in France does not appear to include any changes (such as the introduction of the concepts of combinatorial analysis) likely to account for an improvement in combinatorial thought processes. Besides, with one exception, the participants did not use a mathematical formula to solve the permutation problems. Perhaps the schools promote this type of thinking in an indirect manner. Longeot (1968) provided experimental evidence of such an indirect effect by showing that learning the Cartesian product improves performance on permutation problems. Note also that the presentation of data in table format is much more widespread in the teaching practices of the past few
decades, so much so that table reading and building has become an everyday activity in all disciplines, even as early as elementary school. More generally, one can assume that the development of combinatory thought is promoted by the use of systematic procedures (e.g., in grammar, using a tree to diagram the cases of a rule) and that today’s children are encouraged more to use such procedures.

The scores on the probability, pendulum, and mechanical curves tests did not rise as much. The mechanical curves results are a little surprising. The literature on the Flynn effect would lead one to believe that progress on a spatial test would be greater than on other tests. Indeed, particularly large performance gains have been found on this type of test. However, the mechanical curves test is different from tests like the WISC puzzles, in which substantial improvements have been observed: Although the puzzles involve visual perception skills, mechanical curves require building complex mental representations (note that mechanical curves had the highest loading level on the overall factor in the 13- to 15-year-old group).

Another unexpected result was the one found for the conservation test, in which both an improvement (more participants at Level I in 1996 than in 1967) and a drop (more failures on the weight-volume dissociation) were observed. This test, which has only three items, was not very reliable, and this fact may partially account for the mixed results obtained. Other variables may also have had an impact. Laurendeau-Bendavid (1977) noted that schooling had less effect on conservation (number and area) than on probability quantification. As a consequence, the unexpected conservation test results could also be explained in terms of the respective roles of personal experience and school in the acquisition of the operations. Such an account is in line with the hypothesis that school plays an important role in the Flynn effect. Whatever the case may be, the data collected cannot account for the particular result obtained for the dissociation item: 79% of the 13-to 15-year-olds who failed on this item in 1996 succeeded on volume conservation, and 38% reached the formal level on the LTDS as a whole. Further research is needed to confirm and explain these observations.

If one considers each test separately, the striking finding is the stability of the responses. First, as far as one can tell, the 1996 and 1967 adolescents reasoned in the same way, gave the same arguments, and made the same mistakes. And the item hierarchy for the different operational levels (C, I, Fa, Fb) did not change over the period considered: The interstage indexes were high in both 1996 and 1967. The only changes were on the intrastage indexes, which leads one to suspect certain variations in the item-difficulty order within the operational levels. This information is not sufficient to contend that the formal thought scales measure the same dimensions today as they did 30 years ago, but this hypothesis is not falsified by the data either. The rising performance seems to correspond to genuine progress in the reasoning abilities of adolescents. This point is important for the interpretation of the Flynn effect.

Conclusions

Both studies conducted here showed that Piagetian tests, like traditional intelligence tests, are subject to the Flynn effect. Periodic updating of the norms of these developmental scales is therefore required, just as it is for other tests.

The characteristics of the tasks used in the present study allow one to rule out certain explanations of the Flynn effect—namely, improved guessing and faster test execution speed, as hypothesized by Brand (1987a, 1987b). Furthermore, Study 1 showed that the magnitude of the Flynn effect was the same on the LTDS as on verbal tests (e.g., Verbal scale of the WISC) and smaller on the LTDS than on nonverbal tests. This finding suggests that part of the Flynn effect can be ascribed to gains in visual analysis skills (Flieller et al., 1989; Neisser, 1997), which play a key role in many nonverbal tests but play a minimal role in the LTDS (e.g., the difficulty of the pendulum task clearly lies in the type of reasoning the adolescents have to do and not in their ability to visually perceive the problem data, which are obvious).

Both studies revealed progress on problems involving various kinds of reasoning (e.g., combinatorial problems, proportion problems, control-of-variables problems). Given that the cognitive abilities involved in such reasoning are constituents of intelligence, the Flynn effect can be partially interpreted in terms of a gain in intelligence and a speedup in cognitive development. This interpretation is supported by the fact that the adolescents observed in the 1990s probably solved these problems using the same procedures and the same reasoning processes as the adolescents of the same age observed in the late 1960s. Yet intelligence and cognitive development have many dimensions. The improvement observed here of course only concerns some of these dimensions.

The two studies reported in this article do not allow me to make any statements about the causes of this faster developmental pace. Given the role played by schooling in access to formal thought, the development of academic education is, besides other variables, a serious candidate for the explanation (in his 1996 publication, Flynn himself contended that education may have a major impact in certain countries and during certain periods). In France, where these two studies were conducted, secondary school enrollment grew considerably during the period in question (according to statistics from the French Ministry of Education, the prevalence of middle school attendance at age 15 years rose from 60% in 1967 to 90% in 1993). Note that at the same age, a greater proportion of the adolescents tested in 1967 were in the primary school system (Grades 1 to 8), whereas more of the adolescents in 1993 had gone on to the secondary school system (Grades 6 to 12). The faster pace of cognitive development may therefore be linked in part to the increase in the type of schooling of adolescents. In addition, the adolescents’ parents had also attended school for more years: The average of 9 years in the Longeon and Hormann samples went up to 11 years in the 1993 and 1996 samples (estimates based on national statistics). Given that a study by Lautrey (1980a) showed that children’s operational level depends on the child-rearing practices in the home, which themselves depend on the parents’ education, it is possible that the children’s faster cognitive development is due in part to their more educated parents.

If one assumes that the pace of cognitive development has in fact increased, another conclusion can be drawn from this study. As I mentioned earlier, experiments with operational training of participants instigated many debates in the 1970s. The Geneva School has repeatedly contested the ecological validity of experiments that demonstrate progress after short-term training. Cohort comparisons offer an interesting alternative to the experimental
method for the learning versus development debate. In response to the question Piaget called “the American question,” they suggest that the pace of cognitive development may indeed be rising.

References


Received November 19, 1996
Revision received December 8, 1998
Accepted December 8, 1998