Uncertainty on the measurement of intelligence

A. Martorell Cafranga[a] and J. L. Ayuso Mateos[b]

[a] *Carmen Pardo-Valcarce Foundation.* [b] *Psychiatry Departament. Hospital Universitario de La Princesa. Universidad Autónoma of Madrid. Spain*

La incertidumbre en la medida de la inteligencia

Summary

In recent years, there has been considerable evidence on the phenomenon of intelligence quotient (IQ) gains over time with a gain rate of approximately three IQ points per decade. This phenomenon has been called the Flynn effect. This review article presents the evidence that supports this effect and discusses its implications for the measurement of intelligence. The above mentioned IQ gains over time obviously present serious methodological and theoretical problems in the use of intelligence tests. We review the methodological consequences which the Flynn effect presents in the reliability of the measurement of IQ as well as the methodological consequences which this effect has for epidemiological studies, especially those which focus on the study of the prevalence of mental retardation. Mention is made of the hypothesis that tries to explain these IQ gains, analyzing those which tend to explain these IQ gains as real gains in intelligence, as well as those that propose that these gains are due to other causes rather to real population IQ gains.

Key words: *Intelligence. Neuropsychology. Assesment of intelligence.*

Resumen

En los últimos años se ha puesto de manifiesto que existe un patrón de incremento del cociente intelectual (CI) de 3 puntos por década en diversas poblaciones estudiadas. Esta observación se ha denominado efecto Flynn. Este artículo lleva a cabo una revisión bibliográfica de los datos aportados a propósito de este hallazgo y sus implicaciones en la evaluación de la inteligencia. El incremento antes enunciado plantea serios problemas tanto metodológicos como teóricos para el uso de los tests. En cuanto a los problemas metodológicos se analizan las consecuencias que el efecto Flynn tiene en la capacidad de los tests para dar una medida fiable de inteligencia, así como en lo referente a los estudios epidemiológicos que manejan datos del CI de poblaciones, en especial aquellos en los que se estudia la prevalencia del retraso mental. En lo referente a los problemas teóricos que el efecto Flynn pone de manifiesto, se aborda la cuestión de si los incrementos presentados por Flynn son incrementos reales de inteligencia o si se trata de algún efecto que distorsiona la medida. Se revisan las hipótesis que intentan dar cuenta de estos incrementos, en particular aquellas que apuntan a que existe un verdadero incremento generacional en lo que se refiere a la inteligencia, así como las que intentan explicar el efecto Flynn atribuyendo los incrementos a variables que distorsionan la medida, es decir, variables que incrementan los resultados de los tests pero que no incrementan la inteligencia de las poblaciones.

Palabras clave: *Inteligencia. Neuropsicología. Evaluación de la inteligencia.*

THE FLYNN EFFECT

Since approximately 1918, mean scores of the intelligence tests have been consistently and significantly increasing[1]. From a wide point of view, this phenomenon offers a mean increase of 3 points of IQ per decade, an increase that is almost double in certain specialized measurements. Countries like Holland, where measurements of abstract reasoning are taken every year as part of the military tests, make it easy to observe and measure this increase mean.

In 1994, Herrnstein and Murray baptized this phenomenon in their polemical book *The Bell Curve* with the name of Flynn effect[2]. James R. Flynn was the first to mention this increase when he observed significant increases in the intelligence tests scores over time when studying the results of the Weschler test for his controversy with the geneticist Jensen. Previous investigations had obviated this phenomenon due to the typification of the direct scores in IQ measurements with their stan-

Correspondence:
José Luis Ayuso Mateos
Servicio de Psiquiatría
Hospital Universitario de La Princesa
Diego de León, 62
28006 Madrid (Spain)
E-mail: joseluis.ayuso@uam.es

dardized mean of 100. When Flynn resumed the direct scores and compared them intergenerationally, he surprisingly observed a constant pattern of increase in any one of the intelligence tests or populations studied. Since then, this author has been publishing continually, always rigorously and carefully documenting these increases and proposing many debates to achieve an acceptable explanation of the Flynn effect.

Up to now, 20 countries have been studied (Holland, Belgium, France, Norway, Sweden, Denmark, Germany, Austria, Switzerland, Scotland, North Ireland, England, Canada, the United States, Australia, New Zealand, Israel, Brazil, Japan and China[3]), finding these massive increases of IQ over time in all of them. However, although there have always been differences in the measurements of increases between the different countries, the most interesting differences are those found between the different types of intelligence tests. If we analyze the increase pattern of the last 60 years, the IQ gains have been significantly greater in the tests that are supposedly the purest measurements of intelligence, fluid intelligence tests or measurements of factor *g,* that have also been considered as free measurements of cultural biases. Fluid intelligence tests are those that measure pure mental capacity of problem solving in a given time, independently of the knowledge acquired. On the other hand, we find crystallized intelligence tests that evaluate acquired knowledge, that presumably would vary based on a purer underlying intelligence[4]. The tests that have been most recognized as a measure of crystallized intelligence are the Weschler scales, and as a measure of the fluid intelligence Raven's Progressive Matrices. It is precisely in these latter ones (fluid intelligence tests) where the greatest IQ gains have been found, gains of approximately 20 points per generation (30 years)[1,5 (1)]. The Weschler scales show gains of between 9 and 20 points per generation, the verbal subscale showing an increase of 9 points as a mean. There is not one among the eleven countries that can be compared in which the gains of the «Raven type» tests are not at least twice those of the gains of the «crystallized» scales, finding ratios of 2:1 or 3:1[5]. Even more, subtests such as arithmetic, information and vocabulary do not show increases between generations[5-7]. That is, the subscales and subtests that measure crystallized intelligence within the Weschler tests always show a significantly lower increase than those that correlate with fluid intelligence measurements (understanding, similarities and all the manipulative subscale)[8].

[(1)] The best data available, mainly military tests for young males in Belgium, Holland and Israel, show gains of approximately 20 points per generation (30 years), while Norway, Switzerland and Denmark show gains of about 10 points. Data that are not as rich as the previous ones (Great Britain, Australia and Canada) provide increases of between 12 and 16 points per generation.

WHAT PRACTICAL IMPLICATIONS DOES THE FLYNN EFFECT HAVE IN THE USE OF INTELLIGENCE TESTS?

These results, published for the first time by Flynn in the *Psychological Bulletin*[6], pose questions that are still generating great debates both in regards to the performance of intelligence measurement as well as regarding the underlying theory. In the first place, we will analyze the consequences that the Flynn effect has on the methodology of intelligence tests.

The first practical question arising from Flynn's findings is obvious: if the main utility of the tests is to classify a certain subject in regards to the reference population, and this population is varying, how can the tests' objective be fulfilled? As an example, we use a study of Flynn: the error of the measurement for most of the tests is approximately 5 points, which means that if a subject obtains a result of 100 in a single administration of a test, we have 19 possibilities out of 20 that the real score of the subject will be in some place between 95 and 105. However, if the subject also does not correspond with the generation of the sample used for the test scale (which is quite likely), even if we have the real measurement of his/her IQ in direct score, this could be translated within a range of 90 to 100 or even between 80 to 120, depending on the test type. In summary, the true score of a subject who has obtained an IQ of 100 may be anywhere from 70 to 130[1]. Truly, this is alarming, although, as Flynn states[9], this effect is generally reduced in the professional practice, as the clinical judgment of the evaluator is introduced. The best example to explain how the variability of the IQ influences the accuracy of the tests was provided by Jensen himself. He proposed comparing this phenomenon with a supposed attempt to measure height, based on the measurement of the projected shadows. As the shadow size varies based on the sun position, it is practically impossible to know the real height of a subject. The same presently occurs with the tests: as the scores of the populations vary, it is impossible to know the real position of a certain subject[10].

Thus we cannot forget the Flynn effect when analyzing the different epidemiological studies or those that compare different groups and populations in reference to the intelligence measurements. Good examples of the implications of the Flynn effect on any population analysis are the following studies documented by Flynn: Flynn cites that performed by Vernon in 1982, in which a group of Chinese race North Americans was compared with a group of white North Americans, obtaining results placing the Chinese race North Americans among the intellectual elite in comparison with their co-citizens of the white race. However what was really being compared was a group of Chinese race North Americans with a previous generation of white North Americans, which explains that the intellectual differences were not based on race but rather a generational difference[3,11]. In another case documented by Flynn[6], a group of IQ scores was compared with another group, but of a different

generation, providing incorrect results that gave significant IQ increases to children adopted by families having a higher socioeconomic level when they were compared with their biological mothers. Continuing with this logic, beliefs on intelligence properties can also be questioned: the fact that fluid intelligence decreases with age is a very supported hypothesis. This phenomenon arises when the intelligence scores of adult persons are compared with those of the young people nowadays. However, if we compare the scores of these adults with those of the youth of their generation (those of half a century earlier), the IQ losses would be practically null[12].

Another field closely related with the use of intelligence tests is that of the evaluation of persons with intellectual incapacity[(2)]. Since more than a half century ago, the American Association on Mental Retardation has been the main authority when establishing the diagnostic criteria of intellectual incapacity. In all of its editions of mental retardation definition, ten up to date[13], one of the necessary criteria is an intellectual functioning that is significantly lower than the mean, which in 1983, was consolidated as an IQ of 70 or less[14] [(3)]. What is the justification of the cut-off and how does it affect the classification of the persons with intellectual incapacity? Between 1947 and 1948, the WISC (Weschler Intelligence Scale for Children) was standardized with a representative sample of the North American children of the period. The mean performance was defined as an IQ of 100; performance in the 16th percentile (a standard deviation below the mean) was defined as an IQ of 85; and performance in the 2.27 percentile (two standard deviations below the mean) was defined as an IQ of 70. In other words, an IQ of 70 or lower represents 2.27% of the population that is found in the lower end of the intelligence curve (this is not completely true as most of the intelligence tests only have a «biologically normal» population sample in their scale; that is, they do not include those subjects who suffer brain damages or who present syndromes having a genetic or chromosomic etiology. If they were included, the percentage should increase to 3%[15]. Considering that the question of the development of adaptive behaviors is characteristic in two to three percent of the population, thanks to the mathematics of the bell curve, two standard deviations below the mean would account for 2.27% of a «biologically normal» population[(4)].

That is why there has been an international consensus since 1945 to adopt an IQ score of 70 or less as the cut-off criterion to establish a diagnosis of intellectual incapacity.

Here we run into the first problem: the day after the test ranking, the IQ gains reduce this famous 2.27%. In addition, these gains are sometimes greater among subjects with IQ scores less than 70[1] [(5)].

Thus, between the time when the first WISC was ranked between 1947 and 1948 and the 1972 ranking of the WISC-R, children with an IQ of 70 increased their scores by 8.25 points of IQ[16]. Every year that passes, there are more and more children who exceed the score of 70, so that in 1972, an IQ of 70 only accounted for 0.54% of the end of the curve. Thus, when the WISC-R was published in 1974 with the new standardization, the percentage of those classifiable as subjects with intellectual incapacity dramatically soared overnight to the point of once again including 2.27% of IQ of 70 in the end of the low curve (to immediately go back to including fewer and fewer subjects in the bracket of 70 or less.). In summary, the Weschler tests provide us with a score that supposedly isolates 2.27% of the «biologically normal» population, although this is only momentaneous, since the bell curve begins to shift again to the right, leaving fewer and fewer subjects below a 70 score. This phenomenon obviously has serious repercussions when classifying those with mental retardation: depending on the ranking with which we compare a certain subject, the type of retardation may vary or its existence may even be camouflaged. Thus, we can only know a real IQ of a subject if it forms a part of the sample with which the test has been ranked. Equally, we could use the Flynn effect with objectives different from those of knowing the intellectual functioning: if we want a subject to obtain a diagnosis of intellectual incapacity to thus achieve a better rating of incapacity or some economic benefits, we will apply a recent ranking test, while if what we want is to hide possible intellectual dysfunctions, we will apply a test ranked with past generations.

However, all this also has serious repercussion in regards to the different epidemiological studies of mental retardation. These types of studies are recent in the field on intellectual incapacity, although they are more and more numerous[17]. They mainly try to settle the prevalence and incidence of mental retardation and of its different grades. In most of them, Weschler test measurements are used and the data are analyzed to see the differences in mental retardation rates over time. With that reviewed up to now, it is logical to conclude that the results of these studies cannot provide much information

[(2)] During this article, we will use the term. «intellectual incapacity» and «mental retardation» indistinctly.

[(3)] Three criteria have been established by the AAMR to diagnose mental retardation: significant limitations in intellectual functioning and adaptive behavior with an onset prior to 18 years. It can be stated that the definitions prior to the 9th edition of 1992 did not consider the limitations of adaptive behavior as a diagnostic criterion, only using the measurement of IQ and onset age inferior to 18 years. It was in 1992, with the revolutionary premise that this new definition meant, when the limitations in adapture behavior were included as a necessary criterion, precisely with the intention of eliminating the reductionism and excessive trust in the use of intelligence tests.

[(4)] Flynn tried to trace the data supplied by Weschler for the justification of this 3% of the population, concluding that these data do not seem to exist.

[(5)] Flynn has carefully analyzed the differences in the gains over time between the sample below and IQ of 70 and the rest. These differences are significant, which poses doubts on the measurement of mental retardation.

in regards to the changes in the prevalence and incidence of intellectual incapacity if the Flynn effect is not taken into account.

Then, what can be done to solve all these practical problems presented by the use of the intelligence tests in regards to the Flynn effect? First of all, it should be stated that, in spite of these practical questions, the work behind the elaboration of each intelligence test is admirable in regards to seriousness and magnitude, and that the tests have meant an important thrust for comprehension of intelligence and, more specifically, for the population with mental retardation. However, there are still certain subjects that should be studied in order to continue improving reliability and validity of this type of these. In regards to the control of the Flynn effect, a possible solution would be that of calculating the increase pattern for each test and then of applying it to adjust the measurements based on the date in which they were obtained. Unfortunately, the IQ increase ratio is too variable to be able to establish a constant pattern. For example, when the WISC-III was ranked in 1972, using the ratios of the previous years, gains of about 5 points (0.3 × 17 years) would have been predicted for 1989. The ranking of the WISC-III in 1989 showed these gains of 5 points, however, the ranking of the WISC-III, also with 17 years of difference, showed gains of only 3 points[1]. Who knows if someday, when more data are available on these variations of the IQ, it will be possible to establish a more detailed pattern for each type of test in each population.

Another solution proposed by Flynn is that of abandoning intelligence tests, using behavioral observation scales that measure adaptive behavior[9]. The problem of this option is that these measurements may also be accompanied by some phenomenon similar to that of Flynn, in which case, we would not have accomplished any advancement, but rather, we would have lost all the body of knowledge generated by the intelligence tests. A second problem established by this solution proposed by Flynn is loss of etiological perspective. Low scores on an adaptation skills scale may also be due to a schizophrenic disorder, depression or destructured family setting. With all this, the solution may not be that of replacing intelligence tests with behavior observation measures, but, as the AAMR proposes[13], adding them as one more instrument necessary for any intellectual classification.

In addition, in this section, it is well to mention Jensen's investigations in regards to the possible physiological variables that show correlations with intellectual performance. Jansen's experiments on the electrical response of the brain cortex, on reaction times or on cerebral glucose absorption times[18] may seem to have a promising future. And it may be so in regards to the search for a measurement that is free of cultural biases, however, as occurs with Flynn's behavior observation measurement proposal, who knows if this type of measurements proposed by Jensen are also affected by phenomena similar to the Flynn effect.

Finally, the only remaining solution seems to be that of continually ranking each intelligence test for each population. Flynn proposes a restandardization every seven years to confer minimum respectability to the IQ, although he doubts whether such a periodic ranking would be so costly that is would not be practice[9]. Although it is true that this option is costly and that it may also have side effects that we will review later one, it seems to be a good method to minimize the consequences that the progressive increases in the IQ have on the accuracy of the intelligence tests.

WHAT THEORETICAL IMPLICATIONS DOES THE FLYNN EFFECT HAVE ON THE USE OF THE INTELLIGENCE TESTS?

After this brief review on the consequences that the Flynn effect may have on the use of intelligence tests, we will now focus on the arduous theoretical debate caused by Flynn results on intelligence measurements. In regards to the theoretical doubts posed by the Flynn effect, the first one, obviously, is that referring to whether these IQ increases are real gains of what we understand to be intelligence or if they are explained by another underlying phenomenon. We return to a new practical example of the consequences of the Flynn effect to illustrate this question: if we compare the data obtained by John Raven in two cohorts between 25 and 65 years of age, one measurement in 1942 and the other in 1992, and we compare their direct scores, we conclude that 90% of those born in 1877 are below the percentile 5 of those born in 1967, that is, below an IQ of 75 calculated in 1967[19]. Do 90% of the population born in 1877 really have mental retardation if we observe them with the present rankings? Another clarifying example is one that has already been commented on regarding fluctuations in the percentage of persons with intellectual incapacity, that is, of persons who are below an IQ of 70. If the increase of the IQ scores was really an increase in intelligence, then the number of persons with mental retardation would have really been decreased. However, if this decrease had been real, the 1974 publication of WISC-R, for example, would have alerted all the professionals in the mental retardation setting as it would have once again included 2.27% of the population within the criterion of intellectual incapacity.

Considering all this, is each generation more intelligent than the previous one? Are there underlying factors that account for this phenomenon and that are independent of the intellectual capacity or is it a mixture of both?

Up to now, many hypotheses that we will present in the following have been considered. Normally, they are presented all together, however we believe that there is an important distinction that should be made between those which explain that there are real increases in what we understand as intelligence and those that supporting the fact that the Flynn increases do not reflect real intelligence gains.

Among the hypotheses that state that the IQ increases reflect increases in the intelligence of the population, the following should be stressed:

Improvements in nutrition. It is clear that in the last century, there have been improvements in nutrition, but the relationship that this may have with increases in intelligence is not so clear. Richard Lynn is the main defender of this posture, stating that it only explains the Flynn effect: larger brains produce greater intelligence levels. The IQ increases are thus real increases in intelligence. There are sufficient data that support the correlation existing between malnutrition and low IQ scores, but there is no evidence to the contrary, that is, that better nutrition correlates with greater intelligence. In regards to this point, we bring to mind that gains are found in all the strata of the intelligence curve, not only in those that are below the mean. Furthermore, no correlations have been found between increase in height and increase in IQ And the patterns of increase in intelligence are also not affected by wars or hunger[3].

Socio-economical level and urbanism. These are logical hypotheses since it is obvious that the socio-economic level has increased and that urbanism has improved noticeably, increasing communication and decreasing isolation. However, the existing data poses doubts since the correlations between IQ increase and increase in socio-economic level are not high enough to explain all the Flynn effect[3]. The same occurs with the studies that reflect the urbanization effects of the populations: the calculations of cognitive deficits of the rural zones alone cannot explain the increases found by Flynn[3].

Storfer hypothesis. This is a hypothesis that is similar to that of Lynn, but stresses the eradication of childhood diseases and the improvements of cognitive quality of pre-school and home settings[20]. It can also be understood as a consequence of urbanistic and socio-economic improvements, so that all that mentioned about the two previous hypotheses could be applied to the Storfer premise.

Education. Increase in schooling years and improvement of educational methods and contents are very likely causes that may explain the IQ increases. In regards to the increase of schooling years, they have increased in every one of the 20 countries used in the Flynn studies. Thus, for example, the IQ gains for Denmark have high correlations with the increase in schooling years[21]. However, in Holland, a study of intergenerational pairing to maintain the educational level constant found that this only explained 6.5 % of the gains observed in that country[5].

In regards to the improvement of the educational methods and contents, the results observed also pose problems. A study carried out in Holland in 1982 by Rist[3] demonstrated that when the students educated with the new mathematical educational methods reached military age and were evaluated in the WAIS arithmetic subtests, they not only presented IQ gains in this aspect but also losses.

However, within this section, a new distinction should be made: has the content of the academic curriculum improved or has teaching of abstract skills in problem solving improved? As we explained in the beginning, most of the IQ gains are due to the fluid measurement tests, finding few increases in the crystallized tests. This seems to make it clear that, if there is an educational improvement susceptible to being reflected in IQ gains, this must have occurred in that referring to problem solving. However, how can these educational practices be identified? How can they be measured? Are they really related with fluid intelligence tests?

As can be deduced, the education hypothesis entails complicated premises. To finish, we will cite the 1987 Cahan and Cohen study in which the effects of a school year (controlling age) are compared with those of a chronological year (controlling the schooling years) in 10,000 school children in Jerusalem. As was to be expected, it explained the variance of schooling more than age, but surprisingly, age affected fluid intelligence aptitudes more than crystallized intelligence[22].

New practices of upbringing. With the modernization of the societies, there have been changes in upbringing practices. Parents are interested in the intellectual development of their children and stimulate them in this sense. There are many books on advise for the good up-bringing practices of children, investigations on the subject (Spock, Bowlby, etc.) as well as educational television programs. However, the effect that all of this may have on the IQ development of children is not clear. It is known that pre-school programs such as Head Start do not produce lasting IQ changes. However, we cannot rule out that earlier and more continuous interventions do have lasting effects[22]. For example, when 57 children with a mean age of 4 years who benefited from the Abecedarian Project program in North Carolina were compared with children from a control group, significant differences that did not decrease with time were found. However, this intervention only caused gains of 5 points in IQ, which is far from approaching the gains found by Flynn.

A more technical and visual environment: This hypothesis, proposed by Neisser[22], includes that which has already been stated by different authors: there have been enormous changes in all the countries studied, all of them civilizations that could be considered as more «westernized»[(6)]. There have been changes in ambitions, models, information, leisure activity, etc. Neisser stresses the importance of the visual media: posters, graffiti, movies, television, videogames, computers, etc. and declares that children exposed to these media develop specific visual skills. There is little existing data to analyze this hypothesis, but we include it due to its high intuitive power. The Vincent hypothesis also exists in this sense[23]. This author believes that the complexity of the modern world is responsible and explains the IQ gains. The daily stimuli that we have been subjected to since the industrial revolution are enough to justify these massive increases in intelligence.

[(6)] The samples obtained in China come from urban zones exclusively.

Reduction in lead exposure levels. Our group has recently estimated IQ losses in the populations linked with exposure to lead in their environment[24]. This can cause a reduction up to 2.6 points in IQ for each increase of lead in blood between 10 to 20 µg/dl, assuming a loss of 3.5 points in IQ in levels greater than 20 µg/dl[25]. Measurements of lead in blood obtained for this study in question show lead levels that presently range from 1.7 µg/dl to 15.4 µg/dl, the former corresponding to the mean in the United States and Canada and the latter to Egypt, Morocco and Pakistan. A careful review of these measurements of different countries offers a lead reduction pattern similar to the «westernization» pattern. Thus, we could say that the decrease in lead environmental levels as a consequence of the implementation of governmental measures that favor the use of unleaded gas that has occurred in the developed countries and in many developing countries may explain part of the increases in IQ, although not all of the gains.

We have just reviewed the different hypotheses that seem to indicate that the IQ increases found by Flynn are real increases of intelligence (improvements in nutrition, improvements in socio-economical level and urbanism, Storfer hypothesis, new up-bringing practices, reduction in lead exposure levels). Others, such as more technical and visual environment and a different approach to the effects of education, can be understood as mixed hypotheses, in the sense that they may also reflect other variables that influence the result of the tests but that do not directly affect intelligence. In the following, we will explain the most extreme hypotheses in this aspect, that is, those that would justify the IQ increases but that would not lead to the fact that these increases were real.

The Brand hypothesis. In his hypothesis, Brand[26,27] argues how the increase in permissiveness in societies affects better performance in intelligence tests. The IQ increases are correlated with increases in promiscuity, divorce, lack of religious beliefs, smoking consumption, accidents, crime, since both phenomena are a consequence of this increase in permissiveness. Lack of meticulousness and contempt for rules and lack of fear of the consequences make the subjects answer the tests without wasting time with complicated questions and cause the number of at random answers to increase, which benefits performance, although not intelligence. Raven[28] supplies data that is partially out of keeping with this hypothesis, as he argues that the effect of the at random responses does not increase test performance. Equally, when two generations were compared in the Binet test, we found that it was precisely the second generation that left the most items unanswered[3].

The sophistication of the tests. This hypothesis refers to the increase of the use of tests in the XX century. We have made a literal translation of the name with which it appears in different articles, although perhaps it may be more illustrative to name it something like increase of the presence of tests in society. We are becoming more familiarized with the use of intelligence tests and we have more and more practice in answering their questions. Unquestionable, all this should affect performance; the doubt is to what degree. Flynn cannot contradict this effect, although he tries to minimize it, stating that the IQ increases have not been affected by fluctuations in test popularity[3]. He also states, in this sense, that the subjects who are administered the test repeatedly only offer gains of 5 or 6 points in IQ.

FLYNN'S UNCERTAINTY PRINCIPLE

It is precisely in this last hypothesis where we disagree with that said up to now. After the tests were introduced, society became familiarized with them, learned their rules and the underlying way of thinking and the intelligence tests made an impression on parents, educationalist and students (you only have to analyze the intelligence development programs, which are mostly training in the necessary skills to correctly respond to the tests). That is why the popularity of the tests should not theoretically have an affect, because it is not the number of applications that improve performance, but the way of reasoning proposed; once this form of reasoning appears in a culture, its multiplying effect cannot be stopped, independently of the number of times that the tests are applied. In the same way, the effect of the presence of the tests has nothing to do with a double application of a test to the same subject. This subject has already been born in an environment where the tests are present and its performance will depend on this. The type of reasoning of a test is already known, so that the information obtained on performing a test will mostly be redundant. We will use the sport's world as an example to illustrate this point: a runner of the 100 meters race born in 1992 will obtain much lower times than a runner of the same age, but born 50 years earlier. Even more, both runners have the same physical form. What is this difference in performance due to then? Logically, it is due to the accumulated knowledge on the techniques of athletics. Since the 100 meter race has been established as a test, and since this began to have popularity, each advance in knowledge of training techniques, diets, concentrations, etc., has been accumulating and each new runner does not begin from zero, but rather benefits from this knowledge of a culture that has already known the 100 meter race for one century. The same occurs with any sport: it is sufficient to observe how tennis was played 50 years ago and how it is played now, to see the golf swing of a player 50 years ago and that of the present players, a soccer game then and now, basketball games, etc. However, what does seem to be clear is that these improvements are due to the accumulated knowledge on each sport and not to an improvement in the innate capacities of the athletes. Something may be due to the improvements in nutrition, for example, however it seems clear that these explain the gains less than sport's knowledge. Let's take a tennis player born in 1877 who was among the 10 best and have him play in a league of players born in 1967. What would happen? Although we

equal them in physical form, our 1877 player would be among the 30 worst, but surely due to the accumulated knowledge on the tennis techniques that the 1967 generation has[7].

Thus, seen from this point of view, the argument that performance on a test that has been administered twice to the same subject only improves 5 or 6 points is not valid. A tennis player born in 1967 who plays two consecutive sets does not improve from one set to the other.

A direct example of the test field, and that will serve us later on to explain the differences found within the IQ gains is that of the Raven Progressive Matrices test itself. It is a phenomenon known by all that if we open the Raven booklet to its last pages and see, for example, item D7, it will be complicated to deduce that the correct response is 5. However, if we do the test in its order, we will take little time to correctly answer item D-7 because we have been accumulating knowledge on what the test is asking us for. Even more, Carpenter, Just and Shell[29] have identified the five rules necessary to complete the matrices, that have been acquired during the test: *a)* there is a numerical pattern between the adjacent matrices; *b)* the same value is maintained in the rows but changes in the columns; *c)* one row of a column that is added or subtracted from the other produces the third; *d)* three forms distributed though a row should always be present, and *e)* two values are distributed in a row, and a third value is null. These are, thus, the five rules that we discover during the Raven. In that moment that we decipher one, we accumulate this knowledge, improving performance.

From this new perspective, we will now analyze the different data that Flynn has presented. The most important datum supplied up to now seems to be that of the greater increase rates in the fluid intelligence tests than in those of the crystallized intelligence. The knowledge that is evaluated through the latter has always been present, even before the appearance of intelligence tests. We remind you that the subscales of information, digits and arithmetics have hardly suffered increases (they have even presented losses in some cases: information –0.3; digits +0.1 and arithmetics +0.3)[9]. Appearance of tests has not been able to influence the knowledge that exists on cultural knowledge, numerical memory or mathematics. The same occurs with the vocabulary subtest, with a mean increase of +0.4. However, the rest of the subtests (comprehension, incomplete figures, puzzles, digital symbols, short stories, cubes and similarities) that have presented gains that are 2 to 6 times greater require skills that are not usually present. But there is a piece of data that is even more interesting: in the four tests that hardly show increases, although their items are present in growing difficulty, the order of the response does not affect the result. However, it does affect it in the rest of the subtests[8]. Asking who Gandhi was as item 20 or as item 1 does not affect the result, since the answer is either known or unknown. Repeating the sequence 4-2-7-3-1 before or after a sequence of fewer numbers also does not affect the result as the subject has sufficient memory or does not have it. Proposing a division before or after a subtraction does not affect performance since it depends on previous management that one has of the different mathematical operations. Finally, responding to what the word ominous means in question 33 or before also does not affect it, since one either knows its meaning or does not know it. However, in the rest of the subtests, growing difficulty has an influence, since they require skills that are generally not present, thus the training that the increase in difficulty means influences performance[9]. We believe that this small analysis throws light on the fact that intelligence test gains occur in those reasoning skills that are not generally present and that have been introduced in the appearance of the intelligence tests.

Another piece of data of interest is that supplied by Flynn regarding the fact that there are signs that the increase in IQ gains will reach its end at the finish of the XX century or beginning of XXI, even in the fluid intelligence tests[8]. If we maintain our hypothesis of this new knowledge that the tests have meant, it should not be surprising, that, in fact, at some time it will reach a limit due to saturation of its presence, as has already occurred with the crystallized intelligence tests, approximately in 1948[8].

But, what implications does this new hypothesis have in the use of the tests? We have seen how the introduction of the tests has precisely meant an increase in the skills necessary to fill out intelligence tests, mainly those of fluid intelligence. That is, that the introduction of the tests has caused an alteration in that which they aim to measure. We remember the Heisenberg uncertainty principle: toward the year 1926, the physicist Werner Heisenberg formulated the Uncertainty or Indeterminacy principle, based on an experiment in which he tried to measure two variables: momentum and electron position. As a particle cannot be directly observed, it was necessary to carry out a statistical analysis and then obtain a «likelihood» resulting from the measurement. But the most surprising of this experiment is that when an attempt was made to measure the position, it was no longer possible to determine the momentum, and the same occurred when an attempt was made to measure the

[7] All this makes reference to another debate, approached by Flynn and by the theorists on intelligence, which we want to mention, even though we will not deal with them in this article. That is: what do fluid intelligence tests really measure, pure potential of a subject of performance of this potential?

[8] Perhaps the only questionable subtest in this aspect is that of comprehension, although it could be said that it is precisely comprehension that follows the 4 subtests mentioned as those having less gains.

[9] This is true for all the subtests except for the digital symbols. However, this only has one variable that also has been introduced by the intelligence tests: limited performance time.

momentum, since the position could no longer be determined. Thus, the quantum physicist gave us the knowledge that, on performing a measurement, the behavior of the particles subjected to this measurement changed. Something similar occurs with our hypothesis and the Flynn effect: on using a test as a measurement instrument, we are altering that which we want to measure.

We have made a brief review of the Flynn data, of the practical consequences they have for the use of tests and of the possible solutions, of the theoretical debates that arise and the different explanatory hypotheses. We have finished by focusing on the hypothesis that we consider to be most feasible. We do not want to end without saying that we do not rule out that there may be a conjunction of all the previous causes reviewed among the gains found by Flynn. Undoubtedly, the debate is open and there are increasing attempts to find exhaustive explanatory models of this phenomenon that is known as the Flynn effect.

ACKNOWLEDGES

Work carried out within the FIS G03/061 project.

REFERENCES

1. Flynn JR. The schools: IQ tests, labels, and the word intelligence. En: Carlson JS, Kingma J, Tomic W, editores. Advances in cognition and educational practice. Vol. 5. Conceptual issues in research on intelligence. Londres: JAI Press, 1998; p. 13-42.
2. Herrnstein RJ, Murray C. ART 5. The bell curve: intelligence and class in american life. New York: Free Press, 1994.
3. Flynn JR. IQ gains over time: toward finding the causes. En: Neisser U, editor. The rising curve: long-term gains in IQ and related measures. Washington: American Psychological Association, 1998; p. 25-66.
4. Horn JL. Cognitive diversity: a framework for learning. En: Ackerman PL, Sternberg RJ, Glasser R, editores. Learning and individual differences: Advances in theory and research. New York: Freeman, 1989; p. 61-114.
5. Flynn JR. Massive IQ gains in 14 nations: what IQ tests really measure. Psychol Bull 1987;101:171-91.
6. Flynn JR. The mean of IQ of americans: massive gains 1932 to 1978. Psychol Bull 1984;95:29-51.
7. Weschler D. WISC-III: Weschler Intelligence Scale for children, 3nd ed. Manual. San Antonio: The Psychological Corporation, 1992.
8. Flynn JR. IQ gains, WISC subtests and fluid *g: g* theory and the relevance of Spearman's hypothesis to race. En: Bock GR, Goode JA, Webb K, editores. The nature of intelligence: symposium on the nature of intelligence. Novartis Foundation. Londres: John Wiley and Sons, 2000; p. 202-27.
9. Flynn JR. The hidden history of IQ and special education: can the problems be solved? Psychology, Public Policy, and Law 2000;6:191-8.
10. Jensen AR. Commentary: vehicles of *g*. Psychol Sci 1992; 3:275-8.
11. Flynn JR. Asian americans: achievement beyond IQ. En: Hillsdale NJ, editor. Erlbaum, 1991.
12. Raven JC, Court JH, RavenJ. Manual for Raven's progressive matrices and vocabulary scales. Oxford: Oxford Psychologist Press, 1992.
13. Luckasson R, Borthwick-Duffy S, Buntix WHE, Coulter DL, Craig EM, Reeve A, et al. Mental retardation. Definition, classification and systems of supports, 10th ed. Washington: American Association on Mental Retardation, 2002.
14. Grossman HJ, editor. Classification in mental retardation. Washington: American Association on Mental Deficiency, 1983.
15. Jensen AR. Bias in mental testing. Londres: Methuen, 1980.
16. Flynn JR. Weschler Intelligence Tests: do we really have a criterion of mental retardation? Am J Mental Defic 1985; 90:236-44.
17. Yeargin-Allsopp M, Boyle C. Overview: the epidemiology of neurodevelopmental disorders. Ment Retard Dev Disabil Res Rev 2002;8(3):113-6.
18. Jensen AR. Rising IQ without increasing g? A review of the Milwakee Project: preventing mental retardation in children at risk. Develop Rev 1989;9:234-58.
19. Flynn JR. Searching for Justice: the discovery of IQ gains over time. Am Psychol 1999;54:5-20.
20. Storfer MD. Intelligence and giftedness: the contributions of heredity and early environment. San Francisco: Jossey-Bass, 1990.
21. Teasdale TW, Owen DR. Continued secular increases in intelligence and a stable prevalence of high intelligence levels. Intelligence 1989;13:255-62.
22. Neisser U. Rising Scores on Intelligence Tests. Am Sci, 1997.
23. Vincent KR. On the perfectibility of the human species: evidence using fixed reference groups. Texas Counselling Assoc J 1993;22:60-4.
24. Fewtrell LJ, Prüss-Ústün A, Landrigan P, Ayuso-Mateos JL. Estimating the global burden of disease of mild mental retardation and cardiovascular diseases from environmental lead exposure. Environmental Research 2004;94: 120-33.
25. Fewtrell LJ, Fewtrell L, Prüss A, Landrigan P, Ayuso JL. Burden from environmental lead exposure. En: Murray C, Lopez A, Ezzati M, editores. Quantifying global health risks: the burden of disease attributable to selected risk factors. Geneva: World Health Organization, 2003 (en prensa).
26. Brand CR. Bryter still and bryter? Nature 1987;328: 110.

27. Brand CR. Keeping up with the times Nature 1987;328: 761.
28. Raven J. Methodological problems with the 1992 standardization of the SPM: a response. Personality and Individual Differences 1995;18:443-5.
29. Carpenter PA, Just MA, Shell P. What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. Psychol Rev 1990;97: 404-31.