Orientación y Sociedad ISSN 1515-6877
2015, Volumen N°15– Facultad de Psicología
Universidad Nacional de La Plata

RAVEN'S PROGRESSIVE MATRICES, ARGENTINEAN NORMS FOR AGES 13 TO 18 AND FLYNN EFFECT

Lilia Rossi-Casé, Rosa Neer, Susana Lopetegui, Stella Doná,

Bruno Biganzoli and Ramiro Garzaniti [*]

Abstract

In this study the results from Raven's Standard Progressive Matrices' administration on a sample of 1067 Argentinean students aged 13 to 18 are shown. The number of cases is proportional to the population (INDEC, 2010). For analysis purposes, 3 age groups were defined with equal 2-year intervals. Frequency distributions tend to be normal in all three groups. Observed direct scores increase with age, while variability decreases. Flynn effect is still present, fifteen years after it was observed by this research team in 2000 (Rossi-Casé, Neer & Lopetegui, 2001). In addition, average scores reached after that year show a plateau effect or stability in 15-16 and 17-18 years old groups, and a non-significant decrease in 13-14 year olds ($p<.05$) when compared to 2000 norms. Besides, a change in the difficulty progression between series C and D is observed. All these results are in line with similar research in other countries.

Key words: Raven's Progressive Matrices, norms, 13-18 years old, Flynn effect

---

[*] The authors are Psychologists and Professors at the Psychology School, La Plata National University. Their work is related to the Institute of Psychological Investigations, of the same School. E-mail: rossicase@psico.unlp.edu.ar.

Introduction

In this paper the results of the research project "Test of Raven. General scale. Percentile norms for ages 13-18 for the city of La Plata " are presented, certified by the Faculty of Psychology, National University of La Plata, which was developed between 2012-2015.

Psychological assessment tests comprise the set of instruments –intended for professional psychologists- aimed at arriving at a diagnosis as regards those subjects on which the research rests. These tests provide crucial information for the development of specific intervention strategies for both the diagnosis and prognosis. Direct scores achieved by subjects are transformed to norms that allow comparison with a standardized sample and give a statistical sense to individual performance. However, norms ought to be updated in order to arrive at a correct diagnostic interpretation, i.e, the reference group from which average scores are derived –which are compared against the performance of a subject- should be a proper one. Furthermore, the periodic review of the tests becomes necessary –as regards the new discoveries and hypotheses provided by theories, as well as cultural changes and regional adaptations- for the applied test to be valid and reliable (Casullo, 2009).

A special point should be considered: the average scores on intelligence tests have regularly and significantly increased worldwide. This phenomenon can be observed since the very appearance of psychometric tests. James R. Flynn has summarized the results of research on the subject (Flynn, 1984) and, following this, the steady increase of absolute values in all intelligence tests is called the *Flynn effect*. Hence, as heretofore suggested, psychological testing standards are by no means absolute, universal and permanent. In this regard, Casullo (2009) highlights the need to set up universals for

studying a portion of behaviour through psychological evaluation technique; nonetheless, when setting it up, it ought to respect the particularity of the culture where it will be applied.

This project sought to obtain updated data and compare them with those of the standardization of the years 1964 and 2000, for samples of the same characteristics in order to verify the presence or absence of the Flynn effect in this population.

On the other hand, given the importance of the Raven's Progressive Matrices test (Raven, J.C., Raven y Court, 2003) as an instrument of psychological evaluation and widespread use for different purposes in different areas where psychology is applied, it becomes necessary to update its norms in order to be reliably utilized.

Raven's Progressive Matrices test

The Raven's Test was first published in 1938 by its author, John C. Raven, a student of English psychologist Charles Spearman, who had set forth in 1904 his "Two-Factor Theory or Eclectic Theory". It is a nonverbal test of intellectual and general mental ability. There, the G factor is tested for which it calls into play relationship-elicitation process and correlations on a material in which the variables are not obvious, i.e., new insights should be derived from the information given.

Its administration may be individual or collective, it is self-administered and, in its original version, without time constraints. It's a test of matrices with gaps that should be filled, in which the subject's task is to complete a matrix choosing the correct answer out of six or eight alternatives. It comes in several forms: General Scale, Parallel Scale, Special Coloured Scale and Advanced Scales (Raven, J.C. et al., 2003). The test was revised on several occasions, usually in order to obtain adequate standards or norms for different populations.

The aforementioned elicitive capability is defined as the ability to derive new insights and information based on what is perceived or already known (Raven, J.C. et al., 2003). Elicitation, as a basic cognitive operation, is understood from the three laws in which Spearman splits the knowledge-formation process, known as *noegenetic laws*. The first one, which is called *elicitation relationships law*, states that when facing two or more items everyone tends to establish relationships between them. (Raven, J.C. Raven, J. et al., 2003). The second, which is called *correlates elicitation law*, states that before an item and a relationship, everyone tends to conceive the correlative item. The third, according to which everyone tends to acknowledge itself immediately, as well as the items from its own experience, is thus known as the self-awareness or introspection law. Then, the elicitive behaviour requires a more active perceptual process, rather than analytical or reproductive. It emplies questioning familiar conceptions, as well as developing, understanding and solving problems. Conceptually it would be related to *fluid intelligence*, whereas reproductive capability approaches to *crystallized intelligence* (Cattel, 1968). Fluid intelligence refers to the ability to use abstract reasoning to solve new problems previously untaught (Kaufman, 2009). It is usually assessed by means of abstract analogies whose data are neither outdated nor specific to a single culture. Crystallized intelligence is culturally specific, formal schooling dependent, and subject to the acquired knowledge. Hence it is expected to increase lifelong, whereas the increase of fluid intelligence is not (Kaufman, 2009).

Raven's Test is a test that measures fluid intelligence, therefore we agree with Fernández Liporace, Ongarato, Saavedra and Casullo (2004) when stating that *"this explains such a widespread use of Matrices in research areas,"* since the results provided allow a comparison between populations and / or individuals who *"have been*

exposed to formal and informal educational situations of a rather dissimilar nature" (pp. 50-69).

Due to the confirmation of the intergenerational increase in elicitive capability levels, a "ceiling effect" was achieved in the most capable adolescents and young adults, hence cases where direct test score is the maximum possible became ever more frequent. This introduced the need –after checking it in the British standardization of Raven's Test in 1979 among young people, and in 1992 among adults (Raven, J.C. et al., 2003)- to extend the range of item difficulty in order to restore to the test the discrimination ability which it originally had in these groups of subjects. Thus, in the 1998 revision of the test (Raven, J., Raven and Court, 2003) the Parallel and Plus forms of the General Scale were included.

The General Plus Scale (Raven, J. et al., 2003), which preserves the cyclic format in 60 items of the General Classic Scale, met the need to broaden the range of item difficulty without reducing the test's discrimination power among less able subjects (Raven, J. et al., 2003). Also, in cases where an even greater discrimination between the lower and upper ends of the distribution is needed, there exists the possibility to resort to –respectively- the Coloured and Advanced Scales (Raven, JC et al., 2003).

Simultaneously, a Parallel version developed (Raven, J. et al., 2003). It corresponds generally to the Classical version and item by item in terms of solution strategies and empirical difficulty, and it can be used with existing norms to the General Scale, so that the original and parallel forms are different but equivalents. The creation of this latest version responds to the need to overcome the difficulties generated by the popularity of the test, since due to its being too familiar, some people could train themselves to achieve better results. On this issue, Rossi-Casé, Doná and Garzaniti (in press, 2015) conducted a descriptive correlational pilot study based on a sample of 232 subjects from

the city of La Plata, Argentina. The Pearson correlation coefficient was r = 0.728, with average values and standard deviations similar for both scales.

Flynn effect

It has been noticed, from the widespread use of intelligence tests, that the scores achieved increase regularly and significantly over time *on a world scale*. This indicates that, as years go by, IQ standards become obsolete for the same population. Therefore, at present, direct scores should be higher, and a larger number of problems should be solved, in order to obtain the same, transformed, score as some decades ago. These increases are higher for those tests that measure fluid intelligence, than for those that measure crystallized intelligence (Sundet, Barlaug and Torjussen, 2004). This led to think that the increase is related to the G factor of intelligence and not to specific factors (ie, factors E). The cause of this phenomenon remains unknown and there is insufficient evidence to suggest that this fact reflects a real increase of intelligence, possibly linked to factors such as heterosis, improved living conditions among different populations, better food and nourishment; the expansion of the education system; reducing the size of nuclear families, with a consequent increase of psychological development; the progressive acquisition of certain skills to respond successfully to the tests; further development of schooling and children education; the increasing weight of technology in culture, from video games to the increasingly unfettered access to the media via the Internet, offering other aspects of stimulation (Sundet et al., 2004). For his part, Armstrong and Woodley (2014) claim that there is neurological evidence indicating that the Flynn effect is associated with an increase in brain size that has enhanced the functions of the hippocampus, although do not rule out that another cause for this phenomenon could be the possibility of such tests measuring fluid intelligence and

elicitive ability as the Raven's Progressive Matrices Test, generating some kind of *cognitive scaffolding* that allows the use of skills that are not dependent on factor G to solve problems.

In their initial study, Flynn (1984) established that the magnitude of increase in IQ, initially measured with Wechsler and Stanford-Binet scales, was 0.3 points per year or 3 points per decade. These studies were conducted comparing American samples between 1932 and 1978. Later, the same trend was born out with data from 20 other countries, including Argentina (Rossi Case et al, 2002).

However, recent research shows a decrease in the growth of these scores. This effect was first observed in Scandinavian countries.

Such is the case of longitudinal research conducted in Norway (Sundet et al., 2004) which shows a decrease in scores on a battery of tests administered to 18 year olds, before entering the military service, since the 1950s. The tests used are math and language tests, similar to those subtests of WAIS IV, and a non-verbal test which was constructed similarly to the Raven test. The first two measure crystallized intelligence, while the latter measures fluid intelligence. Teasdale and Owen (2007) studied the data provided by a similar battery in Denmark which, since 1957, is administered to all 18 year olds who enter military service. Of the four tests which compose it, the authors argue that the Letters's Matrices resemble Raven's Progressive Matrices. The authors found that while there was a small increase in scores between 1988 and 1998, they decreased in 2003-2004, even lower than those obtained in 1988 for all battery tests. This was observed in young people of all educational levels. With this in mind, the authors support the assertion that, in this century so far, there has been little evidence about the continuity of the Flynn effect, as defined until now, i.e., as the continued rise in average in intelligence tests.

Meanwhile, Brouwers, Van de Vijver and Van Hemert (2008) did a review of 193 studies on the three Raven scales published between 1944 and 2003, covering 798 samples from 45 countries, the total number of subjects being 244,316. Within that report, Argentina is represented by more than 20 studies published in that period. The authors calculated the correlation between the characteristics of the samples (age, years of schooling) of the countries from which each study came from (Gross Domestic Product [GDP], schooling and illiteracy), and the studies consulted (year of publication). They found out that while both GDP and years of schooling are positively correlated with performance on the test –and hence intercultural differences may be influencing performance- there is a significant negative correlation between GDP and the magnitude of Flynn effect. From this, they conclude that the Flynn effect could have reached its peak in developed Western countries, while countries with lower per capita GDP, still show a steeper Flynn effect. This finding is consistent with Teasdale and Owen (2007), who argue that if the Flynn effect comes to an end in the most developed countries, it would be far from doing so in the rest of the world, but these differences between countries should gradually decline in the future. Flynn (2013) has highlighted that the Netherlands and Finland also show a decrease in the magnitude of the aforementioned effect. However, this trend is not evident in all developed countries. Recent data from Australia are ambiguous about it and the United States and South Korea are clear exceptions to it. He also mentions that some authors attributed this trend to the immigration towards developed countries of populations with a low IQ, but this does not explain the situation in the United States. Flynn ventures the idea that the reason behind the stop of the increase of IQ in some countries could be the fact that they can no longer progress in those areas that are supposed to encourage such increases. Russell (2007) argues that a plateau effect would be expected in cultures that have

optimal living conditions, which would explain why this effect has been expressed first in the Scandinavian countries than in the rest of the world, because they have set up a welfare system that includes the entire population since the end of World War II. Drawing on the Wechsler scales and using lineal regression, this author believes that the United States will reach the plateau effect in 2024. But in considering possible slants in the selection of samples for previous standardization of these scales, he estimates that the plateau could have been reached already in 2004, so that the Flynn effect may no longer be valid in that country.

As regards Argentina, the team conducted an initial confirmation of the Flynn effect in the city of La Plata and its outskirts when updating the Raven test's norms in 2000, and comparing them with the standardization of 1964. This comparison showed a considerable increase in test direct scores for all ages (Flynn and Rossi-Case, 2012; Rossi-Casé, Neer y Lopetegui, 2001, 2002, 2011; Rossi-Casé et al., 2014.). The results were used to check global trends (Flynn and Rossi-Case, 2011).

Method

Participants

The standardization sample settled for 1067 subjects of both sexes, aged between 13 and 18 years old at the time of administration of the test. That range was divided into three equal intervals of two years each. This is because the test performance changes considerably depending on age, which is why the creation of different standards for each age group (Casullo, 2009) is justified.

The sample was stratified considering the variables of age, sex whether the educational institution is either public or private, and high school and university students; from La Plata, Argentina.

Population census data of 2010 (National Institute of Statistics and Census [INDEC], 2010) were taken into account for the extraction of the sample. A two-stage sampling was used, which allowed to select from different conglomerates according to educational institutions and at the same time, in each one of them, randomly choose class groups that would be evaluated depending on the different ages, thus evaluating all subjects that made up each group.

The evaluation was made with a confidence interval of 95% which established that the maximum sampling error would be 3% for this sample size. In Chart 1, the composition of the sample is described, by sex and age. In Chart 2 the percentages of the population and sample composition are compared by age range.

(insertar Tablas 1 y 2)

Instrumental

For the evaluation it was used Test of Raven's Progressive Matrices; General Scale, second edition (Raven, Raven, & Court, 2003). The test consists of 60 problems organized into five series (A, B, C, D and E), twelve items each. In each of the series, items are ordered by increasing difficulty, beginning with simple problems. Each series is with a higher degree of difficulty than before. The increasing difficulty between series is also within each of the five series. Such that the last problem of a series is more difficult than the first one of the immediate subsequent series.

Procedures

The test was administered collectively, in class groups, with the presence of the examiner and without time constraints for execution, to allow evaluation of intellectual capability without involving the speed as a factor to solve the task.

The authorization of Educational Inspectors of La Plata District was requested in a timely manner for the participation of the subjects of the sample. As well as, prior informed parents or tutors of adolescents consent was obtained, using a form which explained the purpose of the investigation and the confidentiality of the data. In the case of students with legal age of majority, at 18 years old in the case of Argentina, the consent was required directly to each of them.

The test was administered in different periods of 2012-2013 and 2014-2015 (November 2012, June, October and November 2013, May and June 2014, March, April and May 2015).

The administration was carried out by professionals and advanced students of the Psychology career, previously trained to standardize the procedure.

To the assignment, the instructions given by the author of the test for collective administration were rigorously followed. The administration was carried out in each of the selected educational institutions in regular class schedule, provided by the authorities and teachers of each institution. Each group did not exceed 36 students.

In forming the database, those protocols that showed a higher than expected discrepancy in the composition of the scores (n = 148) were excluded. Finally a database that includes 1067 protocols was analyzed.

Analysis

With the information obtained a database was created (Excel, Microsoft, 2007) obtaining the corresponding descriptive statistics and percentiles that allowed to elaborate standards for this group.

The results achieved in the present study were compared with the standards achived in 1964 and 2000 for the same age groups in the city of La Plata, Argentina, using the statistic t of Student with a confidence level of 95%.

Data analysis was performed by age group: 13-14 years; 15-16 years and 16-17 years, to compare percentile values with other studies of the same nature. For the same groups they were also compared the scores in the series to estimate the difficulty degree.

Results

The description of the direct scores shows that in the three age groups the range of correct answers varies from 10 to 60. The minimum scores were similar for all ages. The maximum possible score of 60 is reached in the three groups. The results presented in Chart 3 show that, with increasing age of the subjects, performance improvement groups and heterogeneity of their responses decreases.

(insertar Tabla 3)

Comparison through the years, direct punctuation needed for a subject could be located at the 50^{th} percentile, it is presented in Chart 4.

(insertar Tabla 4)

The analysis of the average scores of each Series, by age range, shows that the increasing order of difficulty solving test items is altered in the case of series C and D. In all the age groups studied C series scores were lower than those of the D series.

(insertar Tabla 5)

Results for the 13-14 year age Group

The lowest results of all those comprising the sample are observed in subjects that make up this group. The average score is 44.05 points and the standard deviation is 7.60 points. The range of correct answers was 12 to 60.

The distribution of direct scores necessary for the location of the subject in the percentile scale is shown in Figure 1. The data tend to suit to the normal distribution.

(insertar Figura 1)

Half of the subjects achived 45 points or less. While this represents an increase of 6 points on the norm of 1964, the result is now observed is 3 points lower than that achieved in 2000.

This significant increase in scores on standards 1964 (t (359) = 1.98, p <.05) and normss decrease from 2000, it was observed in all percentile values calculated. These observations can be seen in Figure 2.

(insertar Figura 2)

Results for the 15-16 year age Group

The observed average score is 48.39 points and the standard deviation is 6.75 points. The range of correct responses was 10 to 60.

The distribution of direct scores necessary to locate the subjects in the percentile scale is shown in Figure 3. The data tend to suit to the normal distribution.

(insertar Figura 3)

Half of the subjects achieved at least 50 points. This represents an increase of 9 points on the norm of 1964, and 1 point from the norm achieved in 2000.

This increase in scores on standards 1964 was observed in all percentile values calculated ($t(334) = 1.41$; $p < .05$). Comparison of results with norms of 2000 shows increase of 1 and 2 points for the percentiles 25, 50 and 75, keeping scores for percentiles 10, 90 and 95; and a decrease of 1 point to 5 percentiles 99. These observations can be seen in Figure 4.

(insertar Figura 4)

Results for the age group 17-18 years

The highest results of all who make up the sample are observed in subjects that make up this group. The average score is 49.23 points and standard deviation of 6.39 points. The range of correct responses was 12 to 60.

The distribution of direct scores necessary to locate the subjects in the percentile scale is shown in Figure 5. The data tend to suit to the normal distribution.

(insertar Figura 5)

Half of the subjects achieved 50 points or less. The result observed represents an increase of 10 points compared to the scale of 1964, while remaining equal to the standard achieved in 2000. This significant increase in scores on standards in 1964 ($t(370) = 0.64$; $p < 0.05$) was observed in all percentile values calculated. Regarding the norms of 2000 decreased 3 and 2 points for 5 to 10 percentile values is observed; scores

equal percentiles 25, 50, 90 and 95; and an increase of 1 point in the values of the 75th and 99. These observations can be seen in Figure 6.

(insertar Figura 6)

Discussion and conclusions

The Flynn effect confirmed in 2000 on rules of 1964 (Rossi-case, Neer and Lopetegui, 2001), can't be observed between the scale of 2000 and the current study.

Consistent with recent research in other countries (Brouwers, Van de Vijver and Van Hemert, 2008; Russell, 2007; Sundet et al., 2004; Teasdale and Owen, 2007), the results obtained show an standstill of increased scores direct necessary to achieve the average yield in each age range.

In the current study a slight decrease in the average scores for the age range 13-14 years -which is not significant-, and parity for the other two age groups are born out. Therefore, plateau effect on scores is mentioned.

The Flynn effect shows that the scores have been growing at a rate of roughly one standard deviation generation (Flynn, 1984). It is appropriate to ask whether this increase is indefinite over time or otherwise cease sometime and if so when.

To try to answer this question we should re-examine the hypothesis that once tried to explain the presence of the Flynn effect, such as improvements in nutrition, higher education, the use of new technologies and the predominance of visual images, including other.

Therefore, if conditions ever mentioned to explain the difference have stabilized in recent years, scores, which are their expression, have suffered the same fate.

There is no evidence that there is any other condition that causes a qualitative change in that tenor; then, direct scores should not change now significantly.

Hence, this leads to understanding the plateau effect observed in Argentina and other countries.

Another explanation for this phenomenon could be that the Raven's Progressive Matrices Tests were not measuring what it is supposed, and that other factors were those that affect its resolution, such as formal education and therefore crystallized intelligence.

These factors may be more sensitive to environmental stimulation, which could account not only of the changes that occur over the years, but also the differences between developed countries and those in developing countries.

There is a striking look that emerges from the analysis of the results obtained in this study: it does not accomplish with the progression of difficulty in solving matrices series C and D.

For the three age groups studied, the average direct scores of the C series proved to be lower than those of the D series, opposite result than expected according to the Test of Raven design.

This graduation may influence on the examined subjects, as long as a greater difficulty should be expected when the matrices are happening, and not vice versa, as observed in this study.

They agree with that observed in research conducted by Fernández-Liporace (2004). In both cases suggests the need to further examine these results in order to restructure the C and D series to return to its original design test of increasing difficulty.

CHART 1

Composition of the sample typing Raven Test, 2015
La Plata (Argentina)

Age	Sex		Total cases
	Female	Male	
13 years old	73	135	208
14 years old	68	85	153
Totales del grupo de 13-14 años	141	220	361
15 years old	82	99	181
16 years old	83	72	155
Total group 15-16 years old	165	171	336
17 years old	90	102	192
18 years old	73	105	178
Total group 17-18 years old	163	207	370
Total cases per sex	469	598	1067

Note. Source: Displays, Prepared based test administration General Raven-scale, in 2012-2015, students of both sexes, in La Plata, Argentina (N = 1067). Test de Raven, Carpeta de Evaluación, 2005, Ed. Paidós. Bs. As.

CHART 2

Percentage compared by age range of the population and sample test typing Raven, La Plata (Argentina)

Age Group	Population Census 2010	Sample 2015
13-14 years old	32.44	33.83
15-16 years old	32.75	31.49
17-18 years old	34.81	34.68
Total percentage	100.00	100.00

Note. Source: Population, National Population Census 2010 Ministerio del Interior, República Argentina. Shows, own calculations based on management Raven-Scale Test General, in 2012-2015, students of both sexes, in La Plata, Argentina (N = 1067). Test de Raven, Carpeta de Evaluación, 2005, Ed. Paidós. Bs. As.

CHART 3

Raven test. Average scores, standard deviation and range according to age, 2015, La Plata (Argentina)

Age Group	Average	Standard Deviation	Scores Range
13-14 years old	44.01	7.60	12 a 60
15-16 years old	48.39	6.75	10 a 60
17-18 years old	49.25	6.33	12 a 60

Note*: N* = 1067. Highest score: 60 points.

CHART 4

Comparison of the corresponding direct scores to percentile 50, years 1964, 2000 and 2015, by age range.

Year	Percentil 50		
	13-14 years old	15-16 years old	17-18 years old
1964	29	41	40
2000	48	49	50
2015	45	50	50

Nota*: N* = 1067. Highest score: 60 points.

CHART 5

Average direct scores in each Raven test series, 2015, La Plata (Argentina)

Ages Group	Series A	Series B	Series C	Series D	Series E
13-14	11,30	10,39	8,57	9,02	4,71
15-16	11,57	11,01	9,63	9,88	6,31
17-18	11,65	10,93	9,76	10,22	6,66

Note: (*N*=1067). Raven Test, Carpeta de Evaluación, 2005, Ed. Paidós. Bs. As. Highest score for Series: 12 points

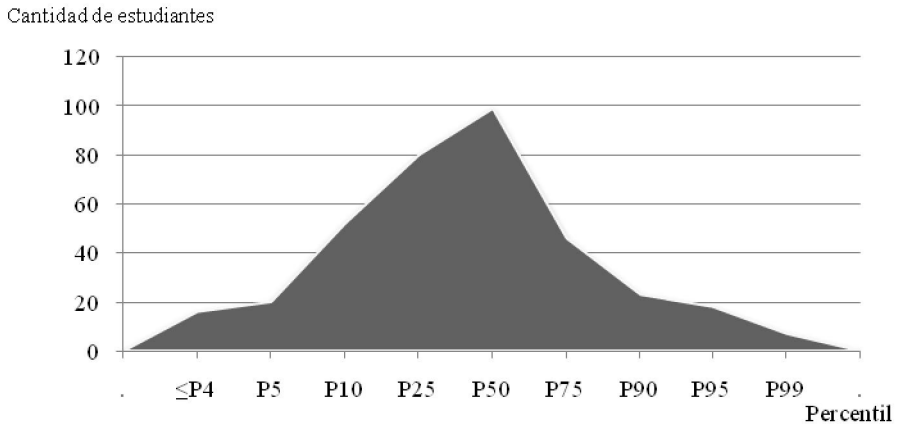

FIGURA 1.
Edad 13-14 años: distribución de la cantidad de estudiantes según el percentil alcanzado (*N*=361)

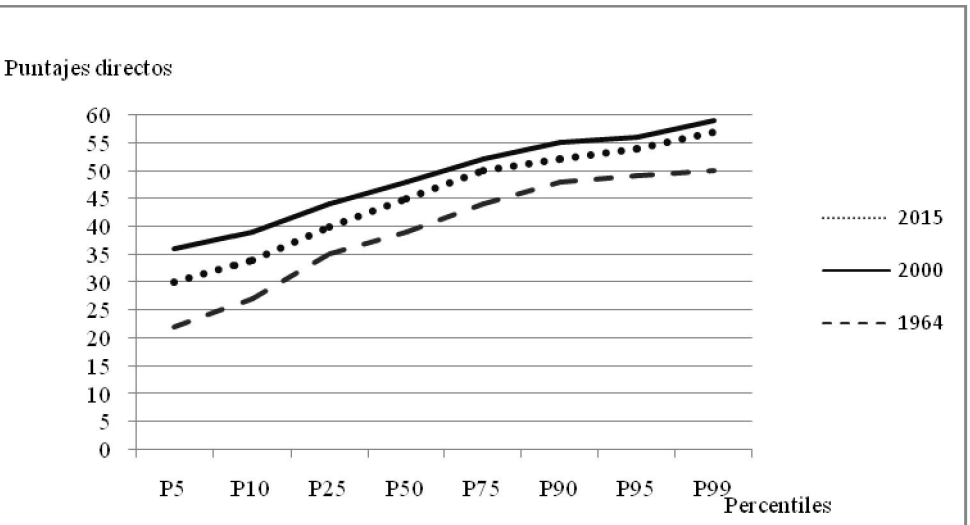

FIGURA 2.
Edades 13-14 años y Efecto Flynn: aumento de las puntuaciones a lo largo del

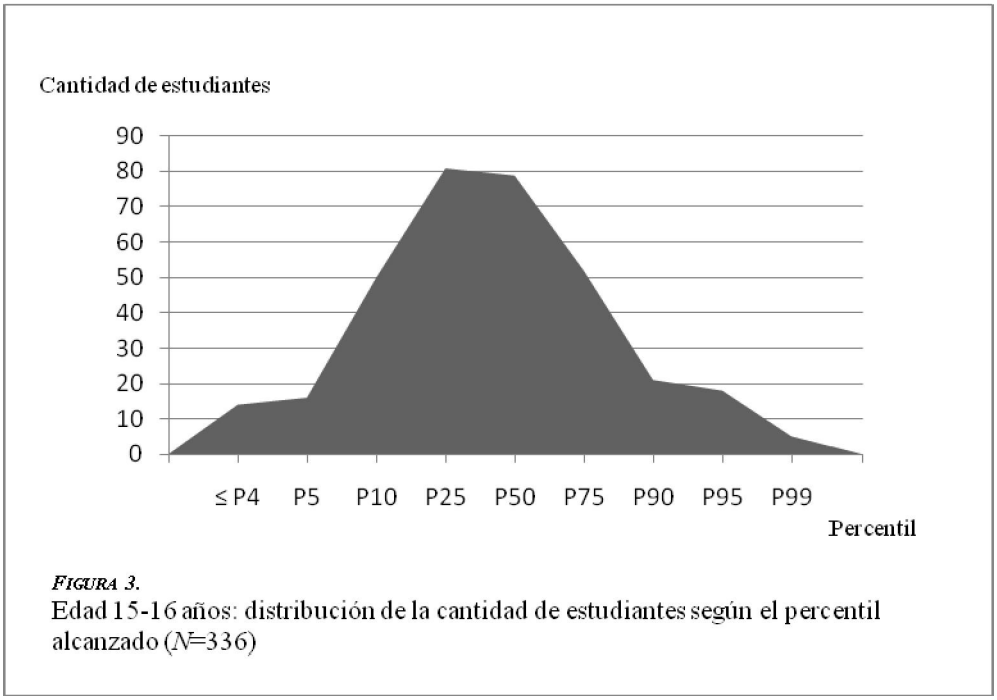

FIGURA 3.
Edad 15-16 años: distribución de la cantidad de estudiantes según el percentil alcanzado (*N*=336)

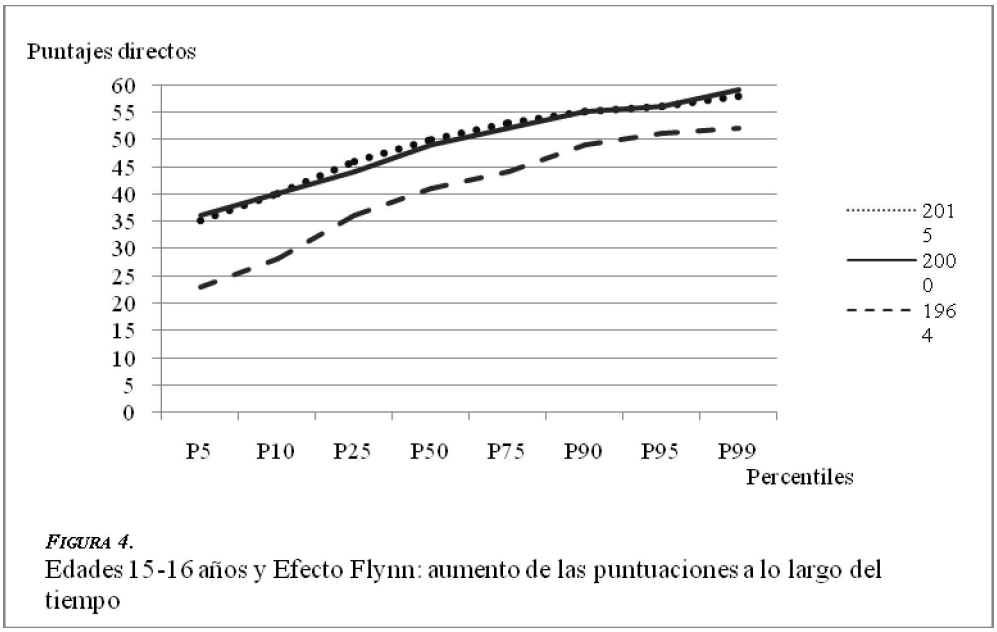

FIGURA 4.
Edades 15-16 años y Efecto Flynn: aumento de las puntuaciones a lo largo del tiempo

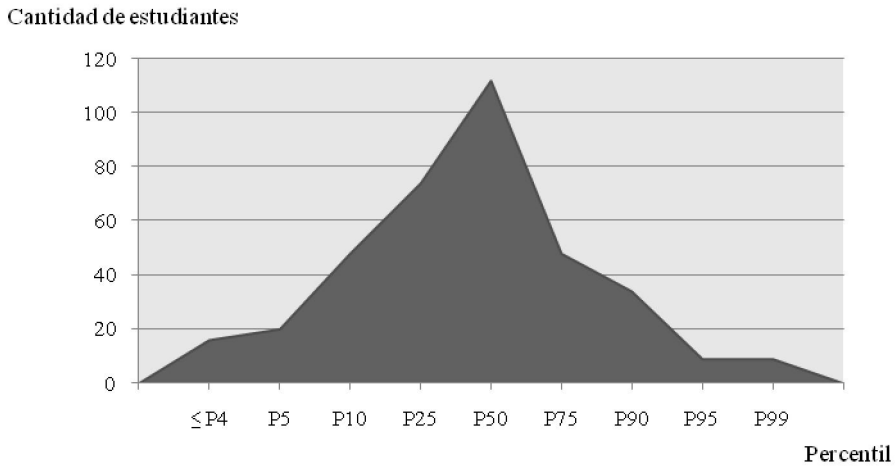

FIGURA 5.
Edad 17-18 años: distribución de la cantidad de estudiantes según el percentil alcanzado (*N*=370)

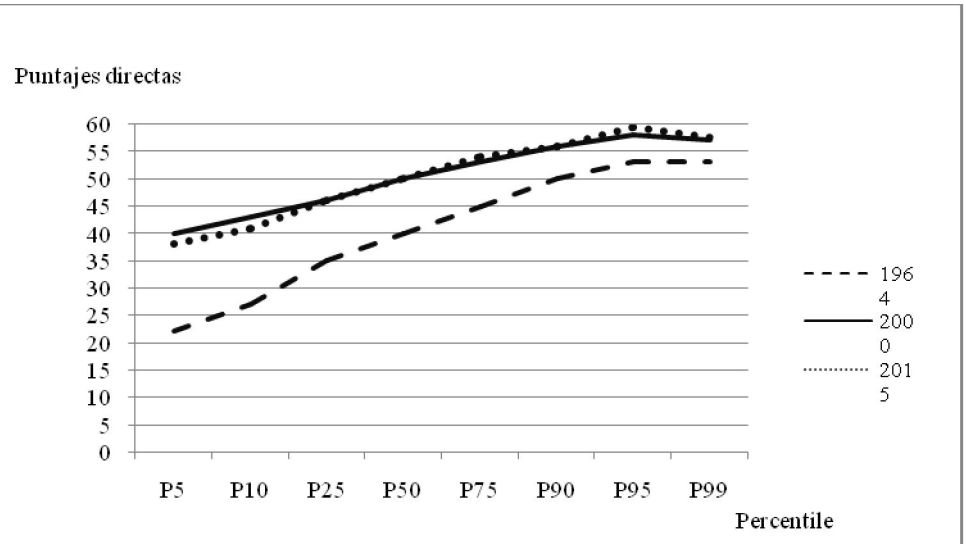

FIGURA 6.