

The Leading Group Effect: Illusionary Declines in Scholastic Standard Scores of Mid-Range Japanese Junior High School Pupils

Kazuo Mori

Institute of Engineering, Tokyo University of Agriculture and Technology, Koganei, Tokyo
and

Akitoshi Uchida

Shinonoi Nishi Junior High School, Nagano, Japan

Abstract

Longitudinal change in the average Z scores for four groups of pupils sorted by quartiles was examined for its stability over three years. The data, collected from 1998 to 2009, was obtained from nine cohorts of Japanese junior high school pupils totaling 1,962 subjects. It showed illusionary declines among the mid-range pupils but improvement among those in the low-range group. These illusionary declines and improvement were found to stem from the fact that left-skewed distributions for the first term examination scores became less skewed for later exams. Such illusionary declines, dubbed the “Leading Group Effect,” may undermine the motivation of middle-range pupils. This illusionary statistical phenomenon should be appropriately explained to teachers and pupils of junior high schools so they won’t be discouraged by illusionary declines in the Z scores.

Keywords Leading Group Effect; Japanese junior high schools; left-skewedness of distribution; standard scores

Acknowledgment: We are indebted to Rebecca Ann Marck for her help in editing the English manuscript. We express our thanks to Prof. Hideki Toyoda of Waseda University, Prof. Tomokazu Haebara of University of Tokyo, Prof. Yasuharu Okamoto of Japan Women’s University, Prof. Tadashi Shibayama of Tohoku University, and Prof. Takehiko Ito of Wako University for their useful comments during the preparation of the draft. Requests for reprints should be sent to Professor Kazuo Mori, Institute of Engineering, Tokyo University of Agriculture and Technology, Koganei, Tokyo 184-8588 or email (kaz-mori@cc.tuat.ac.jp).

To appear in *Research in Education*, **87** (expected publication date May 2012)

Recently there has been increasing emphasis quantitative assessment of educational performance at the individual and institutional as well as national levels. The Organisation for Economic Co-operation and Development (OECD) started the Programme for International Student Assessment (PISA) in 1997, and the first international assessment was done in 2000 with 32 countries, including Japan. The PISA has been conducted every three years and the United Kingdom joined the Programme from 2006. It is important to devise appropriate test items for these educational assessments. It is also necessary to use a proper scoring procedure to compare the test results they produce.

The *Z* score (“*hensachi*”) is a kind of standard score widely used in Japanese junior high schools to assess the scholastic achievement of pupils (Saitoh & Newfields, 2010). It can be easily calculated from the raw scores of any test in any school subject using the following equation:

$$Z = 50 + 10 (X - m) / SD;$$

where *X* = raw score, *m* = mean, *SD* = standard deviation.

The *Z* score has been widely used in Japanese schools because, firstly, the entrance examinations of senior high schools and universities are highly competitive. With essentially only one chance to take an entrance exam per year, it is crucial for a student to know his or her own academic ranking relative to other students of the same age group. Armed with that information, the student can select the highest-ranked school whose entrance exam he or she is likely to pass. The *Z* score has been considered the best way to ascertain where a student stands compared to his or her peers and thus which school to try to enter. Its importance cannot be overstated. Secondly, the *Z* score is robust against the difficulty level of each test. It can easily be compared with the *Z* scores obtained for other school subjects.

If the sample size is considerably large, scores are usually assumed to follow a normal distribution. Thus, the *Z* scores obtained from a sizable group, such as the total number of applicants to high schools in a prefecture, automatically reflect a student’s relative ranking within that group. For example, a *Z* score of 50, which is average, means a ranking almost in the middle of the group, and a *Z* score of 60, which is higher than the average by one *SD*, means a ranking in roughly the top 16% of the group.

However, the normal-curve fitting assumption has often been applied to small groups without appropriate justification. In fact, until recently in Japan it had even been applied to a data set as small as a school class with less than 40 pupils, before the Japanese Ministry of Education changed the grading system for official school records in 2000. Until that year, in Japanese elementary and junior high schools, pupils’ scholastic performance was graded using a five-tier evaluation system. If the class size was 30, pupils were graded in the following proportions: Tier 5, two pupils (7%); Tier 4, seven (24%); Tier 3, twelve (38%); Tier 2, seven (24%); and Tier 1, two (7%). The proportions were calculated by assuming a normal distribution, irrespective of the actual distribution of test scores in the class. Although the five-tier grading system has been changed to a criterion-based one since 2000, *Z* scores have still been widely used as a year-base dataset of pupils, mostly for the purpose of guiding them, their parents and their teachers when they make decisions about the students’ next-level schools.

Although the prevalence of the use of *Z* scores may be limited to Japan, similar standardized scoring systems have been widely used internationally in a variety of areas in

educational practice, such as standardized IQs and the scores used in PISA. The Deviation IQs used in the Wechsler Intelligence Scale for Children (Wechsler, 2003) and the scores used in PISA (OECD, 2009) are standardized scores produced by using the following equations:

$$DIQ = 100 + 15 (X - m) / SD;$$

$$PISA \text{ scores} = 500 + 100 (X - m) / SD;$$

where X = raw score, m = mean scores, SD = standard deviation

In the hypothetical distribution of DIQ, the average is 100 and the standard deviation is 15, while in case of the PISA scores, the average and the standard deviation are 500 and 100, respectively. As for the Z score, the average and the standard deviation are 50 and 10, respectively. Those scores are basically the same standardized scores that can be easily converted from one to the other by means of a simple linear conversion.

In the present study, we examine whether the Z scores obtained from a group of students from a typical Japanese junior high school with six classes in each grade, about 200 pupils in total, accurately reflect the expected scholastic achievements of the pupils. According to the Ministry of Education of Japan (2009), there are about 10,000 municipal junior high schools with 3.6 million pupils in total. Each school has about 360 pupils on average, or 120 pupils in each of the three grades. Therefore, the school from which we obtained our data can be considered a good representative of Japanese junior high schools.

Methodology, Variables, and Data Sets

We examined the longitudinal change in the average Z scores over three years of junior high school for the four quarters of pupils, 0-25 percentiles (U25), 26-50 percentiles (U50), 51-75 percentiles (U75), and 76-100 percentiles (U100), for its stability over three years. It was hypothesized that if all the data were distributed to fit a normal curve, the average Z scores of these four groups would remain constant at about 38, 47, 53, and 62, respectively, regardless of when the assessment was done.

We obtained all the Z scores of term examinations conducted during the academic years of 1998-2009 at a junior high school in Nagano City, Japan. The municipality, with a population of more than 350,000, is located about 200 km northwest of Tokyo and is the capital of Nagano Prefecture, which has a population of about 2 million people. The junior high is a municipal school in a suburb of Nagano City. The socio-economic status of the families of the pupils varies within a narrow middle-class range. All pupils were Japanese natives.

At this school, as at most Japanese junior high schools, term examinations are conducted five times a year: in mid-May, at the end of June, in early October, in late November, and in mid-February (the year-end exam as the Japanese school year starts in April and ends in March.) We used the scores from four term exams, the first mid-term exam in the first year, and the three year-end exams for 1-3 years. Each term examination consisted of a set of achievement tests in the five major school subjects: Japanese language, social studies, mathematics, natural sciences, and English. The five test scores were combined to produce a total score, which was then converted into a Z score. We used only the Z scores for the present analyses.

We obtained data from nine cohorts of pupils between 1998 and 2009, each consisting of about 200 pupils, 1,962 in total. In each cohort, we split the pupils into four quarters according

to their *Z* scores on the first mid-term examination by dividing them at the three quartiles. A considerable number of pupils failed to take the examinations for reasons such as absence on the examination day due to illness, injury, etc., or transfer from the school because of a family move, etc. We omitted all these incomplete data for the purpose of the present analyses, leaving 1,780 complete sets of data in total. It should be noted that the dropout rate for under-achievers in Japanese junior high schools is generally very low mostly because junior high school education is compulsory. Therefore, we assumed that the omission of incomplete data should not affect the longitudinal data analyses for the purpose of the study.

Table 1
Longitudinal change in average *Z* scores of the quartile groups

Quartile	Gender/N		1st yr 1st-mid	1st yr end	2nd yr end	3rd yr end
U100	Boys	Average	59.86	60.87	60.46	61.09
		N = 220 SD	1.70	4.29	4.30	5.59
	Girls	Average	60.19	60.65	59.54	59.76
		N = 250 SD	1.77	3.77	4.24	5.05
U75	Boys	Average	54.94	54.60	54.60	54.69
		N = 213 SD	1.62	4.74	5.20	5.64
	Girls	Average	55.18	53.41	53.71	52.85
		N = 256 SD	1.43	4.89	5.02	5.27
U50	Boys	Average	48.94	47.79	48.82	48.20
		N = 225 SD	2.22	5.34	5.47	5.96
	Girls	Average	48.92	47.07	47.06	46.87
		N = 219 SD	2.46	5.28	5.97	5.81
U25	Boys	Average	36.53	38.34	39.24	39.10
		N = 214 SD	6.42	6.24	6.75	5.65
	Girls	Average	37.69	38.56	38.66	38.44
		N = 183 SD	6.31	5.80	6.13	5.61

Results

The results are shown in Table 1 and Figure 1. As hypothesized from the statistical characteristics of the *Z* scores, the average *Z* scores were basically stable over the three years. But the *Z* scores of the mid-range groups seemed to decline slightly, whereas those of the lowest groups seemed to rise gradually although to a small degree. It also shows that the scores of girls tended to decline more than those of boys. We conducted a three-way ANOVA with the four quartile levels, gender, and examination periods, using the total data of 1,780 pupils to calculate the statistical significances of these points. The results of the ANOVA are summarized in Table 2.

Table 2
Summary of the ANOVA results

Source of Variance	SS	df	MS	F
Quartile (Q)	466599.8	3	155533.3	2325.70 **
Gender (G)	621.7	1	621.7	9.30 **
Q x G	280.7	3	93.6	1.40 ns
Sub (S)	118500.5	1772	66.9	
Exam period (P)	30.3	3	10.1	0.97 ns
G x P	771.5	3	257.2	24.80 **
Q x P	2102.5	9	233.6	22.53 **
Q x G x P	90.8	9	10.1	0.97 ns
S x P	55125.7	5316	10.4	
Total	644123.5	7119		

(ns: not significant, +: $p < .10$, *: $p < .05$, **: $p < .01$)

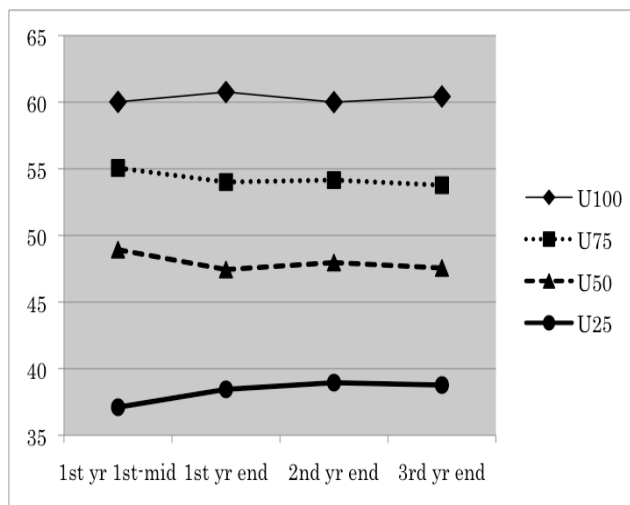


Fig. 1. Longitudinal change in average Z scores for quartile groups

First, as expected, the average Z scores proved to be stable: the main effect of the examination periods was not significant ($F = .97$, ns). However, the Z scores of the two middle-range groups, and those of the lowest groups fluctuated significantly over the three school years: interactions of quartile groups by examination periods were significant ($F = 22.53$, $p < .01$). A *post hoc* interaction analysis was conducted to further examine the interactions of the quartile groups by examination period, showing that the average scores of the two mid-range pupil groups (U50 and U75) declined

significantly from the first mid-term exam to the later exams by 1.50 points for U50 and by 1.29 points for U75 ($MSe = 11.87$, $LSD = .46$, $p < .05$). On the contrary, the average scores of the lowest group (U25) improved gradually and significantly from the first exam (37.11) to the second one (38.45), the third one (38.95), and the last one (38.77). The magnitude of their gains was as large as 1.84 points, reaching statistical significance ($MSe = 11.87$, $LSD = .46$, $p < .05$). As for the average scores of the highest group, although there was a significant difference ($LSD = .46$) between the scores from the second exam (60.76) and the first (60.02) and third (60.00), the differences were within a relatively small range and not consistent.

Table 3
Significances found in the separate ANOVAs

Source of Variance	Total	1998	1999	2000	2001	2002	2003	2004	2005	2006
		-01	-02	-03	-04	-05	-06	-07	-08	-09
Quartile (Q)	**	**	**	**	**	**	**	**	**	**
Gender (G)	**	ns	+	ns	*	ns	ns	ns	ns	*
Q x G	ns	ns	ns	+	ns	ns	ns	ns	ns	ns
Number of pupils	1780	220	175	211	208	209	181	202	197	177
Exam period (P)	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
G x P	**	ns	+	*	**	**	*	ns	**	**
Q x P	**	**	**	**	*	**	**	**	+	**
Q x G x P	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns

ns: not significant, +: $p < .10$, *: $p < .05$, **: $p < .01$

The main effects of gender and the interaction of gender by exam period were also found to be significant ($F = 9.30$, $p < .01$ for the main effect and $F = 24.80$, $p < .01$ for the interaction). A *post-hoc* interaction analysis revealed that the average scores of girls declined statistically (from 50.50 on the first exam to 49.48 on the last one, $MSe = 10.37$, $LSD = .30$, $p < .05$) whereas those of boys rose statistically (from 50.06 on the first exam to 50.77 on the last one, $LSD = .30$, $p < .05$).

No two-way interactions were significant ($F = .97$, ns.). We repeatedly conducted nine separate three-way ANOVAs on each cohort ($Ns = 175-220$), and obtained basically the same results. The significance levels obtained from those ANOVAs are shown in Table 3. It shows that the finding of the long-term decline of the Z scores of the mid-range pupils was consistent across the cohort-wise analyses, while the gender differences were somewhat unstable depending on the cohort sample.

Discussion

Possible effect from the omission of the data

Why did the average Z scores fluctuate so considerably when analyzed group by group? One plausible explanation may come from the fact that there were a considerable number of “drop-outs” from the U25 group data. At the start of the junior high school years, there were 488 pupils. But, since we omitted any incomplete data, by the end there were only 397 pupils, with about 20% of data being lost. On the other hand, the data loss was much smaller in the higher-level groups (U100: 489 → 470, U75: 493 → 469, and U50: 492 → 444). The higher the level, the less incomplete the data were. Probably this was because pupils with lower scholastic levels tended to fail to take the term examinations more frequently than higher-level pupils. If this tendency was indeed true within the U25 group, the number of lower-level pupils would

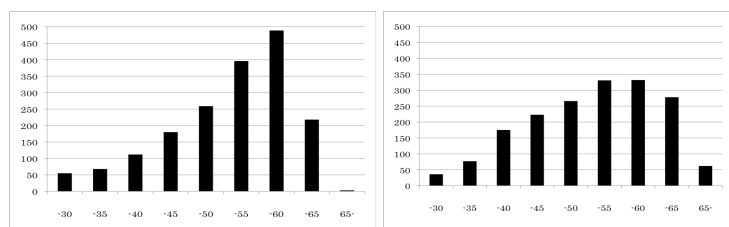
have shrunk gradually whereas the number of students in the upper part of the group who remained to take the later examinations would have been relatively stable.

In order to examine this interpretation, we conducted an additional analysis by re-grouping only the complete data. We selected the data from the 1,780 pupils that had complete records for all the term examinations, and re-classified them into four quartile groups. Then, we re-examined whether the average Z scores of these four groups would change during the three years of their junior high school careers. We obtained basically the same results from this subsequent analysis. The average Z scores of the new U25 group improved from 37.11 to 38.95 by 1.8 points. The two mid-range groups showed the same tendency as the first analysis; the average Z scores of U50 declined from 48.93 to 47.43 by 1.50 points, and those of U75 from 55.06 to 53.77 by 1.29 points. The U100 group data showed the same pattern; their average Z scores improved moderately from 60.02 to 60.76 up by 0.74 point. A three-way ANOVA on the new set of data revealed the same results.

Skewedness of the score distribution: The Leading Group Effect

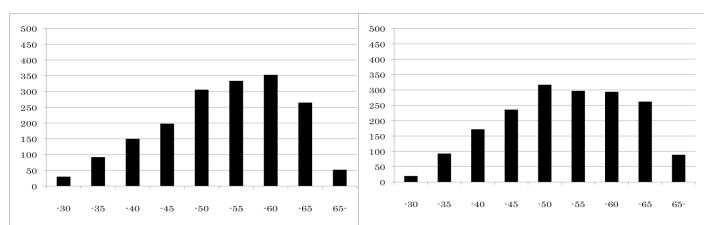
Since the data omission effect did not explain the fluctuation of the average Z scores, we needed to look for other possible causes to explain the phenomenon. Accordingly, we examined the distribution of Z scores over the four examination periods (See Fig. 2). As can easily be seen, the distribution was left-skewed for the first term examination, and gradually became less skewed. We drew the distribution graphs separately for the nine cohorts of the present study and found the same tendency in all of them.

The present study showed that if the distributions are skewed, the mid-range Z scores do not properly reflect the relative rankings among the groups of students. If the distribution of Z scores is left-skewed, as was that of the first mid-term examination of the present study, mid-range pupils get superficially higher scores. Consider this simple example: if a distribution is



(a) 1st mid-term exam at the 1st year

(b) Year end exam at the 1st year



(c) Year end exam at the 2nd year

(d) Year end exam at the 3rd year

left-skewed, such as 20, 55, 55, 55, and 60, a score of 55 means exactly the middle level but the corresponding Z score is 53.7, which is higher than 50. However, if the distribution is changed to a less skewed one, such as 30, 55, 55, 55, 70, the Z score for 55 becomes 51.4. The pupil with a score of 55 remains in the same middle position, but the new Z score shows an illusionary decline.

We named this the “Leading Group Effect,” after the well-known phenomenon of the marathon race in which runners usually form the leading

Fig. 2. Distributions of the Z scores in the four term examinations.

group in the beginning and gradually disperse during the race. Runners in the middle ranks may appear to be in the higher ranks when they are in the leading group, and seem to lose those positions and settle into the middle, although the actual ranks may remain almost the same throughout the race. The left-skewedness we found in the present study seems to be an example of the Leading Group Effect. At the first mid-term examination point there were leading groups, and these groups gradually disappeared in subsequent examinations.

It is worth noting here that there is another type of standard score, McCall's T scores, which are modified Z scores adjusted to fit the normal curve (Guilford & Fruchter, 1978). If T scores are used, no Leading Group Effect will occur. However, the calculation to convert raw scores to T scores is much more complicated than that for Z scores. Therefore, the standard scores widely used in Japan have been predominantly the Z scores.

Should we ignore the differences?

Although the differences we detected among the Z scores reached the statistical significance level, these differences themselves were somewhat superficial. Since the sample size was quite large, with data taken from 1,780 pupils, the power of the statistical test was too strong to detect even a negligible difference (the power for detecting a small effect, $f = .10$, at the significance level of $p < .5$ was .970; Cohen, 1988). The largest difference detected was between the average Z scores of the first and third exams of the U25 pupils; which was 1.84. However, the Cohen's effect size f for this difference was only 0.13. So, should we ignore these small differences despite their statistical significance?

We would agree that educational researchers with a sound knowledge of statistics should ignore such superficial differences. However, we also worry that pupils without enough knowledge of the true meaning of Z scores may well regard these differences as reliable. We strongly suspect that teachers may regard even a small difference between Z scores as a crucial sign of scholastic improvement or decline. If we ignore these differences, we should teach teachers and pupils to ignore them as well. However, teachers and pupils tend to overrate the small differences in Z scores. Therefore, we should not ignore these differences.

Implications for school teachers, educational researchers, and policy makers in education

Educators tend to focus mostly on the raw scores of pupils and the average. Only a small minority concern themselves with the standard deviation. That is why standardized scores, such as Z scores, are used in schools, because even the majority of teachers who usually consider only the average scores automatically take into account the standard deviation. They have been taught that changes in raw test scores do not necessarily mean actual improvement or decline, but those of the Z scores usually reflect real changes. However, even Z scores are not fully reliable if the distribution of test scores is skewed.

Z scores are widely used in Japan because they are regarded as stable measures for scholastic achievement. Teachers and pupils, as well as parents, are concerned about the Z scores throughout the pupils' junior high school lives. They also are concerned about even a small change in the Z scores. In such circumstances, it is worth reporting that there can be an illusionary change stemming from the purely statistical nature of Z scores. Most of the pupils with mid-range scores may have been disappointed with Z scores that declined gradually

throughout their junior high school days despite the fact that their actual scholastic achievement improved and stayed at the same rank among their peers. It may also be true that some improving pupils might have failed to notice their actual progress if it had been canceled out by such superficial downward tendencies.

It should be noted that, although our data were obtained from just one junior high school in a rural district of Japan, a Leading Group Effect can be found in any school as long as the distribution of scholastic scores is left-skewed at the beginning and the skewedness is gradually adjusted in later examinations.

As stated above, T scores are adjusted to fit the normal curve, so the distribution of T scores is not skewed. If T scores are used, no Leading Group Effect will occur. Fortunately, it has become much easier to use T scores instead of Z scores these days thanks to progress in software for assessment in schools.

Why the first term examinations were left-skewed

Junior high school teachers have tended to include easier questions on the first term examination in May so as to produce higher scores for as many pupils as possible. If they get high scores on the first term examination, it might promote self-esteem and provide motivation for studying throughout the pupils' junior high school days. This tendency might be a cause of the left-skewedness of the distribution of the first mid-term examination, because the easier questions would have caused a kind of "ceiling effect" that would cut off the right tail, resulting a left-skewed distribution (See the distribution shown at the top of Figure 2).

There might be still other reasons for the left-skewedness of the first mid-term examination, but the issue is beyond the scope of the present study. The matter should be investigated thoroughly with further studies from various perspectives.

Gender differences

Our data showed that the main effect of gender and the interaction of gender by exam period were significant. The average scores of boys improved statistically, while those of girls declined. Why there was a gender difference is an interesting question. However, we do not have a satisfactory answer, so we will leave this question for future research.

Conclusions

The Z scores are widely believed by teachers, pupils, and parents to properly reflect the relative rankings of pupils compared with their peers. Therefore, if a pupil's Z scores decline, they may think that their scholastic performance has deteriorated. However, the present study reveals that the average Z scores of middle-range pupils showed an illusionary decline caused by the left-skewedness of the distributions of test scores; the Leading Group Effect. Though the magnitude of the decline was as small as 1.5 points or so, the Leading Group Effect may affect the motivation of almost half of the pupils with mid-range scores. Teachers should recognize this illusionary statistical phenomenon and explain it to pupils appropriately so they will not be discouraged by the illusionary decline.

REFERENCES

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, (2nd ed.), Lawrence Erlbaum, Hillsdale, NJ.
- Guilford, J. P. & Fruchter, B. (1978). *Fundamental statistics in psychology and education*, (6th ed.), McGraw-Hill, New York, NY.
- Ministry of Education, Culture, Sports, Science and Technology (2009). *Japan's Education at a Glance 2008*. http://www.mext.go.jp/component/b_menu/other/icsFiles/fieldfile/2009/08/26/1283357_1.pdf
- Organisation for Economic Co-operation and Development (2009). *PISA 2006 Technical Report*. The Organisation for Economic Co-operation and Development, Paris.
- Saitoh, N. & Newfields, T. (2010). Insights in Language Testing: An Interview with Shozo Kuwata --A Pioneer of Standardized Rank Scoring in Japan. *SHIKEN: JALT Testing & Evaluation SIG Newsletter*, **14** (2), 2-5.
- Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). The Psychological Corporation, San Antonio, TX.