Research & Reviews: Journal of Social Sciences

On the Reliability of the Findings of PISA Tests

Shlomo Yitzhaki*
Hadassah Academic College and the Hebrew University, Israel

Research Article

Received date: 11/01/2016
Accepted date: 15/02/2016
Published date: 29/02/2016

***For Correspondence**

Shlömo Yitzhaki, Hadassah Academic College and the Hebrew University, Israel.

E-mail: shlomo.yitzhaki@mail.huji.ac.il

Keywords: Ranking, Mean, Development, OECD, Theoritical condition

ABSTRACT

Knowledge is a hidden variable, and we therefore require a test in order to rank subjects according to their level of knowledge. A test is a battery of questions of varying levels of difficulty. The test results constitute an ordinal variable, since one cannot measure knowledge quantitatively, as one would height or weight. A test can merely rank subjects according to their level of knowledge. It is common practice to rank the success of education systems in various countries according to the average score achieved by students who take a certain international test. An example of such is the PISA test, on which Israel is ranked 29th out of the 33 OECD countries. Averaging is a valid procedure for a quantitative variable, but not for an ordinal variable, the items of which can only be ranked. Since an ordinal variable can be ranked but not averaged, some of the rankings based on averages are unreliable, because one could have devised an alternative test with questions of a different degree of difficulty that would have altered the ranking of the mean scores. This article formulates the theoretical conditions for constituting an alternative test that would alter the ranking of the mean scores, and proceeds to an empirical examination of these cases regarding all possible comparisons between Israel and other OECD countries. The findings show that alternative tests exist that would alter the ranking of Israel's mean scores in relation to half of the OECD member states. This means that in exactly half the comparisons between the OECD countries and Israel, an alternative test exists that would alter the ranking. A further finding indicates that the greater the gap between the mean scores, the less likely one is to find an alternative test that would alter the ranking of the mean scores.

The conclusion to be drawn is that one should attach less importance to ranking according to mean scores.

INTRODUCTION

The PISA (Programme for International Student Assessment) runs a series of tests that constitute an index of the success of Israel's educational system in comparison to those of other countries. The tests are administered by the Organization of Industrialized Nations (OECD) to millions of schoolchildren throughout the developed world at a cost of millions of dollars. Publication of the results are widely reported and has a considerable impact on policy makers and public opinion among the countries that participate in the testing.

In this article I seek to examine the degree of influence that should be attributed to the data pertaining to the ranking of the mean scores achieved by students from various countries that emerge from the PISA test, and to ensure that the degree of influence of the resulting ranking corresponds to the test's degree of reliability. We shall seek to ascertain whether an alternative test exists that relates to the same field of knowledge, the results of which alter the ranking of the mean scores of the various countries. Should such a test indeed exist, this means that we should attach less importance to the results of the PISA test,

since the ranking results depend on how the test is formulated. Through the distribution of the difficulty of the test questions, the composer of the questions deliberately or unwittingly determines the ranking of the mean scores. On the other hand, should we discover that no alternative test exists that alters the ranking of the means, then we may affirm the considerable importance and reliability currently attached to the resulting ranking of mean scores on the PISA tests. We should stress that this is not a matter of statistical significance, which addresses the effect of random error on test results. In this article, we assume that were we to repeat the same test we would obtain the same results, and thus the same ranking of mean scores. One should not treat the "test score" variable as if it reflected a measure such as a centimetre in measuring an individual's height. This is because while altering the distribution of difficulty of the questions in a test does not alter the ranking of the examinees, it does alter the scores themselves and the distance between the examinees' scores, and thus the mean score of the individuals in the group may also change. In the first section of the article we shall explain that since knowledge is a hidden variable, in order to quantify it we must run a test. We then proceed to list the theoretical conditions necessary to compose an alternative test that would alter the ranking of the mean scores. The second section of the article will examine whether one can apply these conditions in the PISA test to Israel's position on the ranking among OECD countries. This procedure is necessary since there may be theoretical conditions that do not exist in empirical reality.

1. The need for a test and effect of the difficulty of test questions on the ranking of mean scores

Knowledge is a hidden variable because it is stored in the subject's mind. The way to expose the extent of a student's knowledge is through testing. The basic assumption is that the higher the examinees' level of knowledge, the better chance they have of answering the question correctly. A second assumption is that the more difficult the question, fewer examinees will answer it correctly. A third assumption is that there is a random component to answering a question, depending on a number of additional factors such as how fatigued or alert the examinee is, his familiarity with the type of question, and so forth. Thus, if we were to repeat the same examination, we should not expect to find that the questions which the examinee answers correctly are identical in both cases. Yet if we assume that the examinees have different levels of knowledge, we should expect to find that if we repeatedly examine the subjects on the same field of knowledge, the random element will decrease and the ranking of the examinees will become more stable. Contrary to height or weight, knowledge is not a quantitative variable that can be measured according to a given unit of measurement such as centimetres or kilograms. The number of questions an examinee answers correctly depends on the level of difficulty of the questions posed. The difficulty of a question is measured by the proportion of those who answer it correctly. The variable of knowledge is therefore an ordinal variable, namely a variable that enables us to rank examinees according to the level of knowledge they demonstrate in the test. We are, however, unable to measure the distance between scores / examinees in given units of measurement. The score that an examinee achieves is an ordinal variable, since the number of correct answers depends on the distribution of the difficulty of the questions. As long as we rank examinees while taking into account that the variable is ordinal, no problem regarding the ranking arises. The problem arises when we employ ranking as if it were a quantitative variable. The rule that applies to ordinal variables may be articulated thus: If the distributions of the scores of two groups of examinees on a test within a certain field intersect, then we can always find an alternative test in the same field that would yield a ranking of mean scores in inverse relation to the ranking that emerged on the present test, provided that the two tests differ solely in the distribution of the difficulty of the questions. But if the cumulative distributions do not intersect, then we are unable to find an alternative test that would alter the ranking of the mean scores. We can prove this theorem mathematically, as demonstrated in several articles on economics, such as Schröder and Yitzhaki 2015 which replicate the propositions developed in financial literature and income distribution and apply them in the area of measurement in education[1]. Here we shall suffice with a simpler demonstration based on inversion of the axes of a cumulative distribution, as in the case of a "guard of honor." While the concept of cumulative distribution is a statistical one, the guard of honor is shown on television whenever a respected figure visits or leaves the country [1]. A guard of honor is formed by placing those who stand in it in a row at an equal distance from one another, with the participants ordered according to height. The first position is occupied by the shortest individual, the second by the second shortest in the group, and at the end of the row (or the beginning at the other end) stands the tallest participant in the group. To ensure that the guard of honor does not depend on its number of participants, it is determined that its length be so that the distance between the participants is 1/N-1, with N representing the number of participants[2]. The practical consequence of this is that instead of employing a ranking that depends on the number of participants who comprise the guard of honor, we convert the ranking into percentiles of the "guard of honor." The guard of honor described above is the cumulative distribution when the axes are interchanged. In other words, if, instead of the regular guard of honor, we interchange the captions of the axes, so that the horizontal axis portrays the height of the participants and the vertical axis represents the cumulative percentage of the population. The following two graphs illustrate the interchanging of the axes between the graph of cumulative distribution and the guard of honor end of the row (or the beginning at the other end) stands the tallest participant in the group. **(Figures 1 and 2).** To demonstrate the problem that arises in using the ranking of groups according to the mean score, let us suppose that we wish to compare two guards of honor, one comprising boys and the other comprising girls. Both guards of honor are standing behind a screen, in order to simulate a hidden variable. Each guard of honor comprises two individuals. The heights of the boys are 160 and 190 centimeters, while the heights of the girls are 170 and 180 centimeters. The

[1]See among others Hadar and Russel (1969), Hanoch and Levy (1969) with regard to Finance, and Atkinson (1970) and Shorrocks (1984) in the area of income distribution.

[2]Since the shortest and tallest participants stand at points 0 and 1, the "distance" between those standing in the guard of honor equals the number of participants minus 1.

test contains only one question. A positive answer is credited with one point, while a negative answer is credited with zero points. Let us suppose that the test question is: who is taller than 185 centimeters? The mean score of the boys is (1 + 0) / 2=0.5, while the mean score of the girls is 0 since all are shorter than 185 centimeters. We would thus conclude that the boys were taller than the girls. If, however, the test question is, who is taller than 165 centimeters, then the mean height of the girls would be 1, whereas the boys' mean would be 0.5. We would thus conclude that the girls were taller than the boys. Therefore, if the guards of the boys and the girls intersect, we would conclude that an alternative test exists that would alter the ranking of the means. If we were to add questions to the test, then the result regarding the relative rankings of the means of the boys and the girls would be dependent on the frequency of "easy" test questions relative to the "difficult" questions. The possibility that the mean scores may be inverted by altering the difficulty distribution of the test questions lies in the ability to divide the guards of the girls and the boys into two parts–up to the point of intersection and above it. In the case of our example, up to the point of intersection, which is the point at which fifty per cent of both the boys and the girls are located, the shortest of the girls is taller than the shortest of the boys, whereas the tallest boy is taller than the tallest girl. And by altering the difficulty of the questions, the examiner can determine on which group the test questions will focus: is the test designed to find geniuses or does it focus on the weaker children? To justify focusing on the weak students, we may argue that the policy of the education system is "no child is left behind." To justify a test that contains mainly difficult questions, we may argue that the test is designed to locate gifted children. If, on the other hand, the cumulative distributions (or the guards of honor) do not intersect, namely if the cumulative percentage of short girls is always greater than the cumulative percentage of short boys, then the mean height of the boys will always be greater than the mean height of the girls, and therefore no alternative test could alter the ranking of the mean scores. If there is only one point of intersection between the groups, the level of difficulty of the test determines which group will have the higher mean score [2,3]. Whenever the groups intersect at least once, we can find two different tests that will rank the mean scores of the boys and girls in a reverse manner. Let us illustrate what happens when there are two points of intersection. To this end, let us add another boy to the guard of honor, so that the heights of the boys (from shortest to tallest) will be 160, 175, and 190 centimetres, and let us assume that the heights of the girls are 170 and 180 centimetres. The test contains only one question. The following **Table 1** shows the possible answers to the question – which is the better group, according to the difficulty of the test. If the number of intersections is greater than one, it becomes more difficult to alter the mean score of a specific group we have chosen by determining the level of the questions. This is because there exist both a more difficult and an easier test that could improve the mean score of the group whose average we seek to improve. In this case, the ranking of groups according to the average score they achieve is random, and is thus meaningless [4]. We assert that comparison of mean scores to the ranking of the groups' success depends on the extent to which the cumulative distributions of the groups' scores intersect. In cases in which there is an intersection, the ranking of the means depends on the distribution of the difficulty of the test questions. The ranking thus depends on the formulator of the test, who may act deliberately (if he or she is sophisticated) or randomly (if he or she is unaware of the significance of their actions).

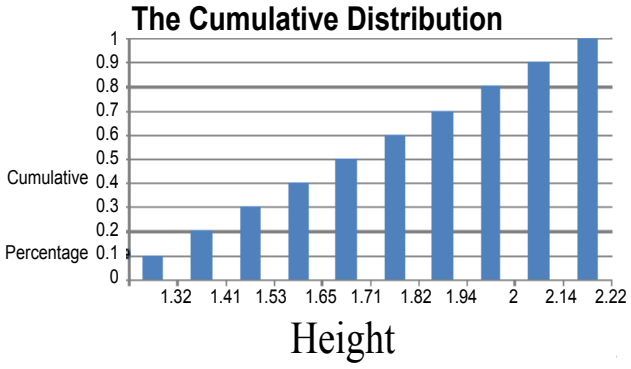

Figure 1. Cumulative Distribution with Height.

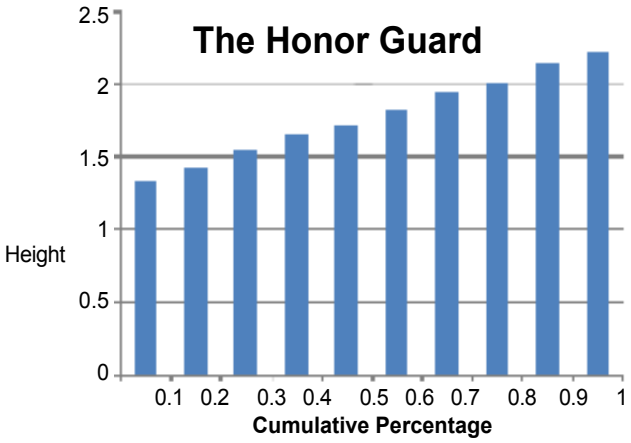

Figure 2. The Cumulative Percentage with Honor Guard.

Table 1. Who has the higher mean score – the boys or the girls?

The test question: Who is taller than	Boys' mean score	Girls' mean score	The higher mean score
165 cm	(1+1+0) / 3 = 2/3	(1+1) / 2 = 1	Girls
172 cm	2/3	½	Boys
176 cm	1/3	½	Girls
181 cm	1/3	0	Boys
185 cm	1/3	0	Boys

NORMAL DISTRIBUTION

To conclude this section, we shall make a further theoretical assertion. Variance represents the spread of scores among the subjects. Assuming that the distribution of the abilities of the groups of subjects (the boys and the girls) is normal, if the variances of the groups are different from each other, then the cumulative distributions will always intersect. If, on the other hand, the variances of the groups are identical, then the cumulative distributions will not intersect.

Proof: The normal distribution is bell shaped. Greater variance means that the bell's amplitude is broader. A broader amplitude indicates a relatively high number of weak and strong examinees. Thus, assuming that the distribution of knowledge is normal, then the cumulative distributions intersects exactly once. This theorem allows us to assert that anyone who assumes that we have a normal distribution of abilities in fact maintains that as long as the variances between the two groups are different, then there will always be an alternative test that alters the ranking of the mean scores.

AN EMPIRICAL COMPARISON: ISRAEL VERSUS THE OTHER OECD COUNTRIES

In this section I present the empirical significance of the theoretical assertions made in the previous section. To this end we shall compare the intersection of the cumulative distributions of Israel on the 2012 PISA tests. Since many countries take part in the PISA tests, so as to spare the reader an overload of details and seeking to make a comparison with the leading nations, the comparison will be made vis-à-vis the other 32 member countries of the OECD, which include Western European countries, the USA, Australia, New Zealand, South Korea and Japan. **Table 2** presents a summary of the instances of an intersection of cumulative distributions of Israel. The furthest left-hand column, titled "Number" denotes the ranking of the 33 OECD members states according to the mean scores achieved on the mathematics test in 2012. The second column from the left, titled "Rank," denotes the location of the country among all the countries that participated in the PISA tests for that year. The third column from the left displays the name of the country. The fourth column denotes the mean score achieved by the students who took the mathematics test. The fifth column indicates the variance of the scores of the subjects in each country. The sixth column displays whether there is an intersection of the cumulative distributions of the country vis-à-vis Israel, and if so, what kind of intersection this is: from above, from below, or multiple intersections. Israel is ranked 29th out of the 33 OECD member states (and 41st out of all the 62 countries that participated in the test). Upon inspecting the "variance" column, we find that the variance of the Israeli students' scores is the highest among the OECD countries[3] Ostensibly, according to the theorem presented in the previous section (regarding normal distribution), if the distributions of skills in mathematics were distributed normally, we should expect the cumulative distribution of the grades of Israeli students to intersect the cumulative distributions of all the OECD countries. Inspection of the intersections of the cumulative distributions reveals that vis-à-vis 16 countries there is no intersection of cumulative distributions, and thus no alternative test exists that would result in a different ranking of mean scores. All the countries with which there is no intersection of distributions have a higher ranking of mean scores than Israel. On the other hand, with regard to the four countries ranked lower than Israel there does exist an alternative test that would lower Israel's ranking on mean scores. It should be noted that with regard to the 12 countries located above Israel on the ranking of mean scores with which the cumulative distributions intersect, this does not mean that there exists an alternative test that would rank Israel at 17, since the comparison is made separately with each country. For some countries the alternative test would be easier, while for other countries the alternative test would be more difficult. We did not determine the maximum number of ranks that Israel could climb, since an algorithm that could perform this calculation has yet to be developed.

Table 2. Results of the 2012 PISA test and the instance of an intersection with the cumulative distribution of Israel.

Number	Rank	Country	Mean score in mathematics	Variance	Intersection
1	5	South Korea	554	9291	None
2	7	Japan	536	8273	None
3	9	Switzerland	531	8344	None
4	10	Holland	523	8006	None
5	11	Estonia	521	6105	None
6	12	Finland	519	6783	None
7	13	Canada	518	7412	None
8	14	Poland	518	7680	None
9	15	Belgium	515	9934	None

[3]We admit that this is the reason for choosing Israel to demonstrate our point.

10	16	Germany	514	8821	None
11	18	Austria	506	8065	Multiple
12	19	Australia	504	8809	None
13	20	Ireland	501	6724	Multiple
14	21	Slovenia	501	7949	Multiple
15	22	Denmark	500	6299	Above
16	23	New Zealand	500	9379	None
17	24	Czech Republic	499	8547	None
18	25	France	495	9029	None
19	26	Great Britain	494	8429	Multiple
20	27	Iceland	493	7861	Above
21	29	Luxembourg	490	8502	None
22	30	Norway	489	7677	Multiple
23	31	Portugal	487	8361	Multiple
24	32	Italy	485	8123	Multiple
25	33	Spain	484	7228	Multiple
26	35	Slovakia	482	9685	None
27	36	USA	481	7622	Multiple
28	38	Sweden	478	7896	Multiple
29	41	Israel	466	10411	--------
30	42	Greece	453	7081	Above
31	44	Turkey	448	7769	Multiple
32	50	Chile	423	6037	Above
33	52	Mexico	413	4970	Above

SUMMARY OF FINDINGS

As mentioned above, the variable “knowledge” or “ability” is a hidden variable to which no natural unit of measurement applies. To ascertain level of knowledge we require a test. The test comprises a number of questions intended to reveal the examinee’s level of knowledge. Yet since the score achieved on the test depends on the distribution of the questions’ level of difficulty, the easier the questions, the higher will be the score. All we can expect of statistical methods is to rank examinees according to level of knowledge. Despite this limitation, experts and economists tend to calculate mean success of examinees in various groupings, and some apply regressions designed to find correlations between success in studies and success in other fields. In this paper we have formulated rules that enable us to ensure that no alternative test, with a different distribution of difficulty of questions exists, which would alter the ranking of the mean scores. Upon conducting an empirical examination of the comparison of findings indicating the success of examinees in the 2012 PISA tests in mathematics, we found that in 50 percent of the cases checked a valid alternative test exists that would enable us to alter Israel’s ranking vis-à-vis other countries. Choosing an OECD country at random and flipping a fair coin would give a similar degree of accuracy. The advantage of flipping a coin over testing is that flipping a coin is cheaper than examining millions of students. We further found that the smaller the differences in mean achievement, the higher the likelihood of finding an alternative test that would alter Israel’s ranking. This conclusion can only be determined by an exam. Finally, it should be noted that if we are dealing with an ordinal variable, the use of the regression method of identifying a correlation between examinees’ knowledge and economic variables, as is common practice among economists, is similarly flawed. That is, it is possible that an alternative test exists that would alter the sign of the regression coefficient. For a demonstration, Schröder and Yitzhaki (2015).

ACKNOWLEDGEMENT

I am grateful to Dvir Miller of the Hadassah Academic College who assisted in analyzing the data. My thanks to Nili Gefen, Ruth Klinov, Ashik Movshovitz, Ruth Ottolenghi and Guy Yitzhaki, who read a previous draft of this article and whose comments helped me to improve the presentation of the findings.

REFERENCES

1. AB Atkinson. On the measurement of inequality, Journal of Economic Theory. 1970; 2: 244-263.
2. J Hadar and WR Russel. Rules for Ordering Uncertain Prospects. American Economic Review. 1969;59: 25-34.
3. G Hanoch and H Levy. The Efficiency Analysis of Choices Involving Risk. Review of Economic Studies. 1969;36:335-346.
4. C Schröder and S Yitzhaki. Revisiting the evidence for a cardinal treatment of ordinal variables. DIW working paper. 2015;772.
5. AF Shorrocks. Ranking Income Distributions, Economica. 1984;50: 3-17.