סדרת ניירות עבודה

WORKING PAPER SERIES

No. 37 מס'

עמידות המסקנות המבוססות על הציון הממוצע במבחני ידע בינלאומיים
The Robustness of Conclusions Based on TIMSS Mean Grades

שלמה יצחקי*, אפרת סופר ועדנה שכטמן****

Edna Schechtman, Efrat Soffer** and Shlomo Yitzhaki***

July 2008 יולי

* הלשכה המרכזית לסטטיסטיקה –הסטטיסטיקן הממשלתי
** אוניברסיטת בן-גוריון בנגב

*Central Bureau of Statistics – Director of the Bureau
**Ben-Gurion University

L H B E N

Published by the Central Bureau of Statistics, 66 Kanfe Nesharim St.,

Corner Bachi St., P.O.B 34525, Jerusalem 91342, Israel

Tel. 972-2-6592666; Fax: 72-2-6521340

Internet Site: www.cbs.gov.il

E-Mail: info@cbs.gov.il

Abstract

Globalization and computerization enable the wide use of comparisons of average grades achieved in standardized exams by students from different nations and the rankings of the nations, based on the average grades of their students. The results are reported in front pages of newspapers. This paper suggests a method for examining the robustness of those results and illustrates it by analyzing the results of five countries reported in TIMSS. In more than sixty five percent of the comparisons examined it is found that there exists an alternative legitimate exam that will reverse the ranking of the countries.

KEY WORDS: **Ability, Measurement, TIMSS**

1. Introduction

There exists an ample of econometric evidence that links education and especially the quality of education to personal economic affluence and to economic growth. Robert Barro (2001) in a paper entitled "Human Capital and Growth" which also summarizes his book (1997) writes:

> "Data on students' scores on internationally comparable examinations in science, mathematics and reading were used to measure the quality of schooling. Scores on science tests have a particularly strong positive relation with growth. Given the quality of education, as represented by the test scores, the quantity of schooling, measured by average years of attainment of adult males at the secondary and higher levels is still positively related to subsequent growth. However, the effect of school quality is quantitatively much more important." (AER, May 2001, pp-16-17).

As a result of this and other evidence, results of average test grades that present comparisons between countries, districts, schools, etc. are used for evaluation of different teaching methods, rewards for identifying better schools and evaluating different methods of teaching. Numerical results are viewed as hard evidence that is difficult to argue with. It shows the exact effect in a seemingly precise way. The aim of this paper is to present additional evidence that one should not accept average grades at face value, and it is important to distinguish between the way we treat average values of grades, even if they are derived from a valid exam, from the treatment of other quantitative variables such as the average income or average height.

There is a major difference that distinguishes between height and grades, which serves as the base for our argument. Height can be **directly** measured while ability is a latent variable. To measure ability – a questionnaire is composed and grades are determined as a monotonic increasing function of the number of correct answers. For a given distribution of abilities in the population, the difficulty distribution of the questions in the exam determines whether grade is a monotonic concave function or a monotonic convex function of ability. The type of function that connects between grades and ability cannot affect the ranking of individuals' abilities but it may affect the ranking of averages of grades among groups. The necessary and sufficient conditions for the existence of an alternative legitimate exam that can change the ranking of average grades are stated and proved in Yitzhaki and Eisenstaedt (2003). They were illustrated in Schechtman and Yitzhaki (2006) which dealt with comparisons of the success of ethnic and other groups in matriculation exams in mathematics in Israel. It shows that in about 40 percent of the cases examined, there exists an alternative exam that can reverse the ranking of the groups. This paper adds two dimensions to the earlier papers. It applies the methodology to the arena of the prestigious international testing and comparisons and more importantly, it adds empirical analysis and interpretation that enable one not only to determine that an alternative exam **exists** but also to characterize the alternative exam that can reverse the results. This adds an operational aspect to the methodology. The presentation is restricted to the analysis of average grades of five countries: Australia, Bulgaria, Israel, Romania and USA, using six different types of exams, taken from TIMSS, the Trends in International Mathematics and Science Study. The restriction to five countries is intended to simplify the illustration of the methodology without bombarding the reader with too many results. Of the 60 comparisons examined, 41 comparisons turned out to be inconclusive. That is, one can find an alternative test that will reverse the ranking of average grades of the groups. We note that because we illustrate our point using data from five countries only (already leading to sixty comparisons), the above summary figures should be taken as illustrative only. The main conclusion is that a careful examination of the results according to the suggested methodology is called for before reaching a conclusion concerning the ranking of the countries.

The structure of the paper is the following: In Section 2 we detail the methodology. Section 3 describes the data, while the main results are presented in

Section 4. Section 5 offers conclusions and suggestions for further research that is called for.

2. The methodology

In this section we describe the theoretical arguments. Detailed mathematical proofs can be found in Yitzhaki and Eisenstaedt (2003). The theory is based on several underlying assumptions. First, it is assumed that ability, which is a latent variable, is one dimensional.[1] As a result of this assumption the exam is a legitimate one for evaluating the ability of individuals. By a legitimate exam it is meant that if one individual has higher ability than another then the scores of the higher ability individual cannot be lower than the scores of the lower ability one. Second, it is assumed that there exists an increasing monotone relationship between performance (grade) and ability. We will use the term a legitimate exam whenever the capability of the exam to grade individuals according to one-dimensional ability is not challenged.[2]

Formally, let the probability of a correct response for a given question (a "hit") be p(a, d), where ***a*** is the subject's ability and ***d*** is the difficulty of the task.[3] For convenience, we assume that both ***a*** and ***d*** are continuous. Neither ***d*** nor ***a*** are observed. The purpose of the questions is to rank the members of the population according to ability, where we assume that the more able the subject, the higher his probability of success (for a given ***d***) and the harder the task, the lower the probability of success; formally, ***p*** is increasing in ***a*** and decreasing in ***d***.[4] We assume

---

[1] A multi-dimensional ability is much more complicated to handle because it may make the grades sensitive to the different type of abilities. This problem is known in the literature as the Simpson's Paradox. (See Wainer, 1986a,1986b, 1994; Terwilliger and Schield , 2004). The Simpson's paradox does not apply to uni-dimensional ability.

[2] Real-life exams may include several legitimate exams, each one of the them converts one-dimensional ability into grades.

[3] Lord (1980, p. 12) refers to this function as an *item response function*.

[4] This assumption is known as the "monotonicity assumption." Additional assumptions that could be imposed on equation (1) below are local independence and local homogeneity (see Ellis and Wollenberg, 1993), but these additions are not relevant to our main argument.

that the grade of a subject with ability ***a*** on a test of difficulty ***d*** is also affected by "white noise" errors, that is:

(1) $$g(a,d) = p(a, d) + e$$

where ***g*** is the observed grade, p(a,d) is the expected grade, and ***e*** is a white noise error with mean zero (with respect to ability and with respect to ***d***), and with the regular assumptions that the errors are not correlated with ability or with ***d***. Note that item-response models assume that the difficulty is fixed, and hence look at models of the type: grade = h(a) + e. We deal with a more general situation, where the grade is a function of the ability as well as the difficulty distribution of the questions, ***d***.

As the investigator is the one who sets up the test, which is composed of several questions, he also controls its difficulty distribution (intentionally or unintentionally). The score (and the probability of success) in a test with n questions, administered to a subject with ability ***a***, is:

(2) $$S(a, d) = \frac{1}{n} \sum_{i=1}^{n} g(a, d_i),$$

(***d*** is a vector whose components are d_i). Equation (2) states that a subject's observed grade is the average of n random variables which represent the grades on the individual items of the exam. However, these random variables are not statistically independent — they are all affected by ***a***, the subject's ability. S(a, d) is a random variable which represents the score of an individual with ability ***a*** in a test with difficulty ***d***. In constructing an exam it is assumed that, as a rule, tests are constructed in accordance with certain accepted rules-of-thumb which serve to ensure that components will have discriminatory power without being redundant. It therefore follows that any attempt to re-structure a test will have to abide by similar rules. In the case of TIMSS the policy concerning the difficulty distribution is stated on the website. It says:

> "Items are reviewed by an international Science and Mathematics Item Review Committee and field-tested in most of the participating

> countries. Results from the field test are used to evaluate item difficulty, how well items discriminate between high- and low-performing students, the effectiveness of distracters in multiple-choice items, scoring suitability and reliability for constructed-response items, and evidence of bias towards or against individual countries or in favor of boys or girls. As a result of this review, replacement items are selected for inclusion in the assessment."

As was stated above, we define a legitimate test as a test composed of questions with different difficulties which follows the above mentioned rules, so that the probability of answering each question is a non-decreasing function of the ability, ***a***. In other words, a test is meant to be legitimate if its capability of ranking individuals according to their ability is not challenged.

The full characterization of the probability of scoring a hit requires additional assumptions on the interaction between ability and difficulty: does an increase in ability have a greater (smaller) effect on the probability of scoring hits as the task grows more difficult? Since we do not know the answer, and wish to keep the presentation as simple as possible, the following functional form is assumed:[5]

$$g(a, d) = p(a,d) + e = h(x)+e= h(a - d)+e \quad (3)$$

where $x = a - d$ measures the difficulty of a task for a subject whose ability is ***a***, and conversely, the ability of a subject to correctly answer a question of difficulty ***d***. The assumption we make is that the derivative obeys $h'( ) > 0$, which means that the harder the task, the lower the probability of scoring a hit, and that the higher

[5] This assumption is typical to many models that assume unidimensionality of the response function (see Ellis and Wollenberg, 1993; Rasch, 1966; and Brogden, 1977). However, our main argument is still valid even if a general function p(a,d) is assumed.

the subject's ability, the better his chances of scoring a hit. Finally, we assume that there exist x_{max} and x_{min} such that:[6]

(4) $$h(x) = 0 \quad \text{for} \quad x \le x_{min} \qquad \text{and} \qquad h(x) = 1 \quad \text{for} \quad x \ge x_{max} .$$

Assumption (4) means that one can always compose a question that no one will ever answer correctly, and another that will always be answered correctly. This assumption eliminates the possibility of the probability of success being a constant that is independent of the task's difficulty.

It is worth noting that although the problem of ranking groups versus ranking individuals is presented in a stochastic model (i.e., an Item Response Theory model, (Lord, 1980)), the basic problem — being able to affect the ranking of groups — may exist even in a deterministic model in which $h(x) = 0$ for $z > x$ and $h(x) = 1$ for $z \le x$ for some z. Clearly, the stochastic case is the common one. Therefore, we concentrate on the stochastic case.

The objective of this paper is to show that in some cases there **exists** another legitimate test that will result in different ranking of groups' averages. We intend to show that in the case of a test that examines one attribute, ranking groups differs from ranking individuals; the latter is not sensitive to the test's difficulty distribution. On the other hand, the ranking of groups may, under certain circumstances, be sensitive to the test's difficulty distribution, and hence can be affected by the difficulty distribution of the questions in the questionnaire, d. The aim of this section is to identify such cases. The identification of such cases sheds some light on the robustness of the ranking reported by the exam. An additional outcome is that it tells us where along the ability distribution is the strength (or the weakness) of one group with respect to the other, so that it can direct countries about the strategies that can help them pass other countries. We emphasize that we **do not** intend to imply that the TIMSS developers may not be doing a good job in

[6] These are assumptions of convenience; the conclusions reached here are not affected by allowing a "guessing parameter" to affect the item response function (see Lord, 1980, p. 12).

item analyses, thus some countries maybe artificially better than others. We only challenge the robustness of ranking, based on means.

The first proposition summarizes the conclusion for the trivial case of ranking individuals.

PROPOSITION 1 (Yitzhaki, S. and M. Eisenstaedt (2003), adjusted for the non-deterministic case).

Individuals' ranking within a group cannot be altered by changing d.

Proof: Let two individuals have abilities $a_1 > a_2$. We will show that $E\{S(a_1, d)\} > E\{S(a_2, d)\}$ for all d. According to equation (1),

$$E\{ S(a_1, d)\} - E\{S(a_2, d)\} = \frac{1}{n}\sum[h(a_1 - d_i) - h(a_2 - d_i)] > 0.$$

The non-negativity of the terms in the square brackets is obtained from the assumption that $h'( ) > 0$.

QED

Group ranking, being more complex, needs an example. Take two groups of equal size, "blues" and "greens", where $a_1^b \le a_2^b \le ... \le a_m^b$ and $a_1^g \le a_2^g \le ... \le a_m^g$ denote blues' and greens' abilities, respectively. Denote the cumulative distribution function of group b at ability a by $F_b(a) = \frac{1}{m} \sum I(a_i^b)$, where $I(a_i^b)$ is 1 when a_i^b is less than or equal to a and zero otherwise. That is, $F_b(a)$ is the proportion of subjects of group b with abilities less than or equal to a. Because F() is unobservable, we will use the empirical cumulative distribution function instead, that is $F_b(S) = F_b(S(a, d)) = \frac{1}{m}\sum I(S_i)$. Clearly, F(S) is a function of the abilities in the group, and the difficulty distribution of the exam. The ranking of the groups is determined by the difference in average scores achieved in the test, as follows:

$$(5) \qquad E\{\Delta R\} = \sum_{j=1}^{m}[E\{S(a_j^b, d)\} - E\{S(a_j^g, d)\}].$$

Can the sign of $E\{\Delta R\}$ be changed by changing d?

PROPOSITION 2

Assuming that (5) is used to rank groups, and that equations (2) and (3) hold, then a necessary and sufficient condition for the impossibility of changing the sign of $E\{\Delta R\}$ by an alternative selection of the vector d is that $F_b(a)$ and $F_g(a)$ do not intersect.

Proof: Begin with a test in which all questions are equally difficult, $d_1 = d_2 = \ldots = d_n = d_c$. In this case, it suffices to prove the proposition with a test composed of one question.

Suppose the distributions intersect only once, at a_0 That is, $F_g(a) > F_b(a)$ for $a \leq a_0$ and $F_g(a) < F_b(a)$ for $a > a_0$. If so, one can choose d_c such that $a_0 - d_c < x_{max}$. Now, all the subjects whose $a \geq x_{max} + d_c$ will score a hit with probability one, and since $1 - F_g(x_{min} + d_c) < 1 - F_b(x_{min} + d_c)$ there will be fewer greens than blues among them. As for the rest, since $F_g(a) > F_b(a)$, the blue with the poorest ability has better chances of scoring a hit than does the green with the poorest ability, the blue second in rank is more likely to score a hit than is the green second in rank, and so on. Blues will therefore perform better than greens in this test.

To change the groups' ranking it suffices to choose d_c such that $a_0 - d_c < x_{min}$. Now, only $1 - F_g(x_{min} + d_c) > 1 - F_b(x_{min} + d_c)$ will score a hit. Scanning from best to worst: the best green has a higher probability of scoring a hit than the best blue, the second-best green has a better chance than the second-best blue, and so on. This proves that if the distributions of two groups intersect, one can switch their rankings. If the distributions do not intersect, then for any d_c chosen by the investigator, if the lowest ranking member of one group has a higher (lower) chance of scoring a hit than does the lowest ranking member of the other group, then the same can be said of the rest of the population.

QED

Note that the conditions as stated in proposition 2 are quite common. If the unobserved ability distributions are assumed to be normal, it is sufficient for the variances of the two groups to differ to cause an intersection of the ability distributions. This means that assuming normal distributions of abilities means that when the variances are not equal, then the examiner can cause rank reversal of average scores of groups simply by changing the difficulty distribution. This property holds regardless of the values of the expected abilities in the groups (i.e., the means of the normal distributions). However, the difficulty of finding the alternative test that can reverse the order of the mean scores of the groups is a function of the difference in expected abilities. To see this, note that the density function of the normal distribution with the higher variance intersects the one with the lower variance twice. This implies that the higher variance distribution has a higher proportion of both excellent and bad students.[7] Hence, by an appropriate choice of difficulty distribution the examiner can affect the ranking of the means.

The condition of non-intersecting distributions is identical to the condition that the groups can be stochastically ordered (Lehmann, 1955; Spencer, 1983a,b), or, to use the term used in economics, that the distributions can be ranked according to First Degree Stochastic Dominance criterion (FSD).[8] The intuitive explanation to this result is the following: the difficulty distribution of the exam determines the function h as a function of a. In a range of abilities with no question to distinguish between the examinees, $\frac{\partial h}{\partial a} = 0$, and the greater the number of questions with the

[7] If the cumulative distributions intersect once, then the density functions intersect twice.

[8] The first-degree stochastic dominance (FSD) criterion states the following: assume two distributions, $F_b(a)$ and $F_g(a)$, and let $S(a)$ be any function with $S'() \geq 0$. Then $E_b\{S(a)\} \geq E_g\{S(a)\}$, where E_b is the expected value for all functions $S()$ if and only if $F_g(a) \geq F_b(a)$ for all a. See, among others, Copeland and Weston (1983), pp. 92–93; Huang and Litzenberger (1988), pp. 40–43; Saposnik (1981). Levy (2006) offers a comprehensive survey.

same difficulty, the greater $\frac{\partial h}{\partial a}$. Hence, by selecting the difficulty distribution the examiner selects the function h(a) from the set of functions h(), with $\frac{\partial h}{\partial a} \geq 0$.[9]

The proof of Proposition 2 relies on three assumptions: (i) the groups are equal in size, (ii) the test consists of one question, and (iii) the distributions intersect only once. The proposition can easily be extended without these assumptions.

An important property of Proposition 2 which will come into play later is that whether or not the distributions intersect does not depend on d. This is so because the subject's ranking is not sensitive to the test's difficulty distribution (see proof of Proposition 1). Since the cumulative distribution results from the individuals' ranking, changing the difficulty distribution cannot change the order in which the cumulative distributions are ranked. However, for a test to reveal more than gross clumping of performance levels, it is important that the empirical cumulative distributions be strictly increasing; a test that is too easy or too difficult may obscure finer degrees of differentiation in the subjects' abilities.

The similarity between the conditions of means' score reversal and First order Stochastic Dominance (FSD) rules enables us to borrow additional results from FSD. As is well known from the theory of stochastic dominance, the method results in a partial order among the distributions . That is, whenever one compares two groups, there are three possible outcomes: either one group dominates the other, or that it is impossible to find dominance. Similar cases should hold in comparing group averages. Our purpose is to identify the cases where it is impossible to rank the groups.

---

[9] This description holds for a questionnaire with an infinite number of questions. In a questionnaire with a given number of questions, the examiner selects a restricted function.

3. Data description

TIMSS, the Trends in International Mathematics and Science Study, is designed to help countries all over the world improve student learning in mathematics and science. It collects educational achievement data at the fourth and eighth grades to provide information about trends in performance over time together with extensive background information to address concerns about the quantity, quality, and content of instruction. Approximately 50 countries from all over the world participate in TIMSS.

The sources of our data are the records of all students of the 8th grade who participated in the TIMSS's mathematics exam, year 2003, which were downloaded from the site of International Association for the Evaluation of Educational Achievement (IEA).[10]

The TIMSS 2003 eighth-grade assessment contained 194 items in mathematics. Between one-third and two-fifths of the items at each grade level were in constructed-response format, requiring students to generate and write their own answers. The remaining questions used a multiple-choice format. In scoring the items, correct answers to most questions were worth one point. However, responses to some constructed-response questions (particularly those requiring extended responses) were evaluated for partial credit, with a fully correct answer being awarded two points. The total number of score points available for analysis thus somewhat exceeds the number of items.

The types of exams that are examined in this paper are: Mathematics overall, Number, Geometry, Data, Measurement, and Algebra. A detailed description is presented in Appendix A. Our working assumption is that each exam is testing a property that obeys our basic assumptions, namely: (1) it is

---

However, since if there is an intersection a questionnaire with one question only is sufficient to change the ranking, we can ignore the above restriction.

[10] Internet site:http://isc.bc.edu/timss2003.html

composed of a one-dimensional ability that can be described by a (latent) scalar, and (2) if Edna has higher ability than Shlomo then the probability that she can answer any question correctly is higher than Shlomo's probability. In other words: in our analyses we assume that each type of questionnaire evaluates the ability of students in one and only one dimension. (Otherwise, other issues like Simpson's paradox may arise, which add additional reason to suspect the robustness of the conclusion). However, it is clear that this assumption is violated in "Mathematics overall" and we assume it in order to be able to ignore the Simpson's paradox (Wainer, 1986a,b, 1994).[11] The data used in this paper contains the raw grades of the 6 types of exams for 5 countries: Australia, Bulgaria, Israel, Romania and USA.

4. Results

The number of possible comparisons in this research is very large because for each type of exam it includes all possible comparisons between any two countries. (That is, the number of exams * number of combinations of two countries). In order to illustrate the results we chose four examples which will be dealt with in detail. Then, we will summarize our findings in a table. In the four examples we will report the following:

1. A table of sample sizes, averages and standard deviations, maximum and minimum grades – which composes the set of descriptive statistics.
2. A graph which represents the difference between the two (empirical) cumulative distribution functions as a function of the grade. **A** graph which intersects the horizontal axis represents a case where the empirical distributions intersect, which implies that there exists another legitimate exam that will reverse the ranking of the average grades. Note that the group for which the empirical cumulative distribution is higher is the one with lower grades.

[11] Wainer and Brown (2004) point out two paradoxes that can be attributed to conclusions with respect to groups. The point of this paper can be considered as an additional difficulty.

Because the empirical cumulative distributions are composed of the data at hand (random sample), crossing between cumulative distributions can occur in the sample, but it need not reflect the relationship between the populations. For this purpose, we present the results of Kolmogorov-Smirnoff test (KS) for testing the equality between two cumulative distribution functions. We point out that KS does not test for intersections. To the best of our knowledge there is no test available for intersection of two cumulative distribution functions. Tests for intersection of two Absolute Concentration Curves (ACC) was recently proposed (Schechtman et al, 2008), but some further research is needed before it can be used for testing the intersection of two cumulative distribution functions. We use KS in this paper as a first screening only. That is, if the empirical cdf's intersect, but KS is not significant, then we conclude that the two cdf's do not differ significantly and therefore the means are not significantly different and there is no point in looking at the ranking. There is a need for further (theoretical) research to develop a formal test for the intersection of two cdf's, but it's beyond the scope of this paper. We illustrate the idea by plots.

Example A: Mathematics overall - Australia vs. Israel

In this example we compare Australia versus Israel, ignoring the fact that "overall" is composed of several attributes.

Table 1. Descriptive statistics of grades of Australia and Israel in mathematics overall

Group	Australia	Israel
Sample size	4791	4318
Average	497.57	495.10
Standard deviation	76.27	81.17
Maximum grade	785.76	741.69
Minimum grade	233.95	231.5

Based on the averages, as shown in table 1, Australia achieved higher average and one can conclude that they are better students in "mathematics overall". We can see that the worst student is an Israeli, while the best student is an Australian which is an additional indication that the Australians are the better group. Let us now check whether there is an alternative test that will reverse the ranking of the groups. In order to do so we plot the difference between the cumulative distribution functions, as illustrated in Figure 1. Let us remind the reader that the higher the cumulative distribution- the lower are the grades in the group.

Figure 1. The vertical difference between the cumulative distribution function for Australia and Israel in mathematics overall

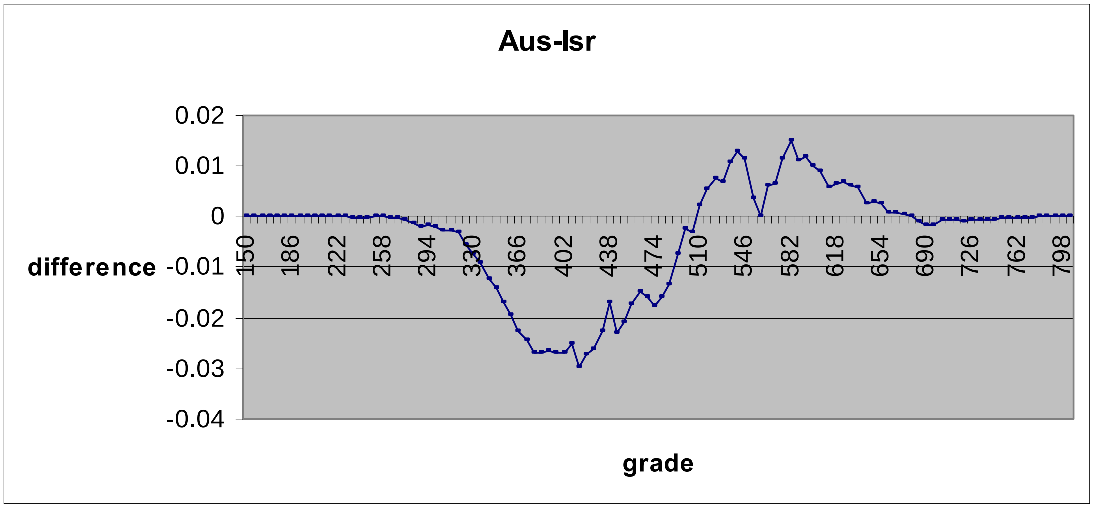


Figure 1 presents the vertical difference between the empirical cumulative distributions of Australia and Israel. In general, this curve has several important features that will be useful in evaluating the possibility of finding an alternative exam that will reverse the ranking of average scores between the countries. The features are:

1. The range on the horizontal axis is the range of the grades in the two countries.
2. The height of the curve at grade **g** represents the difference in ability between Australia and Israel up to that grade level. If the curve is negative (positive) at **g**, it means that there are relatively more Australians (Israelis) with a higher grade than **g**. To see this note that $F_A(g) - F_I(g) < 0$ implies $\{1-F_A(g)\} > \{1-F_I(g)\}$ where $F(g)=P(\text{grade} \le g)$.
3. The total area enclosed between the curve and the horizontal axis (positive contribution when above the axis, negative contribution when below it) is equal to the difference in the means of the distributions of Israel minus

Australia. [12] As can be seen the negative area in Figure 1 is larger than the positive area, reflecting that Australia has a higher average score than Israel.

4. The slope of the curve represents difference between the density functions of Australia and Israel. A positive slope implies relatively more Australians than Israelis at that grade level while a negative one implies relatively more Israelis than Australians. [13]

Given these properties we now turn to the investigation of the implications of Figure 1 on the possibility of finding the alternative test that will reverse the ranking of average grades.

The grades vary in the range (231, 786). In this range we observe a negative part of the difference between the cumulative distribution functions (260-510) and a positive one (510-680). From property 2 we gather that at any grade level **g** in the range (260-510) there is a higher proportion of students with higher ability than **g** among the Australians than among the Israelis. On the other hand, at each grade **g** between (510-680), there is a higher proportion of Israelis with higher ability than **g** than among the Australians. An alternative exam with more questions that can be answered by those in grade levels between (510-680) and fewer questions that can be answered by grade levels (260-510) will improve average grade of Israel relative to Australia. To reverse the ranking of average grades we should continue changing the questionnaire until the target is achieved. The higher (lower) the curve the higher the advantage (the disadvantage) of Israel. Therefore, to achieve our target in a minimum number of changes in the questionnaire we should add questions around the peak of the curve (around 582) and delete questions at the minimum of the curve (around 410).

---

[12] To see this note that the expected score, μ, is equal to $\mu = \int_0^\infty [1 - F(x)]dx$. (This result can be derived by integration by parts of the regular definition of expected value). Therefore, given two distributions, A and I, one gets $\mu_A - \mu_I = \int_0^\infty \{F_I(x) - F_A(x)\}dx.$

[13] To see this note that the curve is $F_A(x) - F_I(x)$ so that the derivative is equal to $f_A(x) - f_I(x)$.

Note that if the curve does not cross the horizontal axis, for example if it would be negative all over the range, then at every grade level **g**, there will be a higher proportion of Australians with higher ability than Israelis, and one would not be able to change the ranking of average scores. The relatively large range (510-680) in which there is an advantage to Israel is an indication that finding the alternative test would not be too difficult.

The slope of the curve tells us where the Australians (Israelis) are located. Whenever the curve is increasing, (e.g., 402-540) then by property 4 there is a higher proportion of Australians at that range. As can be depicted from Figure 1, there are relatively more Israelis than Australians in the ranges [260-402] and [582-610] and relatively more Australians between those two ranges. This explains why the difficulty distribution of the questions in the questionnaire matters and can affect the ranking of the countries.

Another indication to support our conclusion is that the variance among the Israeli students is higher than the variance among the Australians, meaning that in Israel there are more relatively weak and relatively good students, relative to the Australians, making the difficulty distribution of the exam crucial in determining which country will achieve higher average grade.

Conclusion: It is possible to derive a test with a different distribution of difficulty of the questions, in which the ranking of the groups according to the averages will be reversed. A test which will reverse the order of ranking will consist of more hard and fewer easy questions. The p-value of the KS test is 0.022 therefore the hypothesis that the cumulative distributions are identical is rejected. Hence, our conclusions are not derived as a result of random errors.

Example B: Geometry - Israel vs. USA

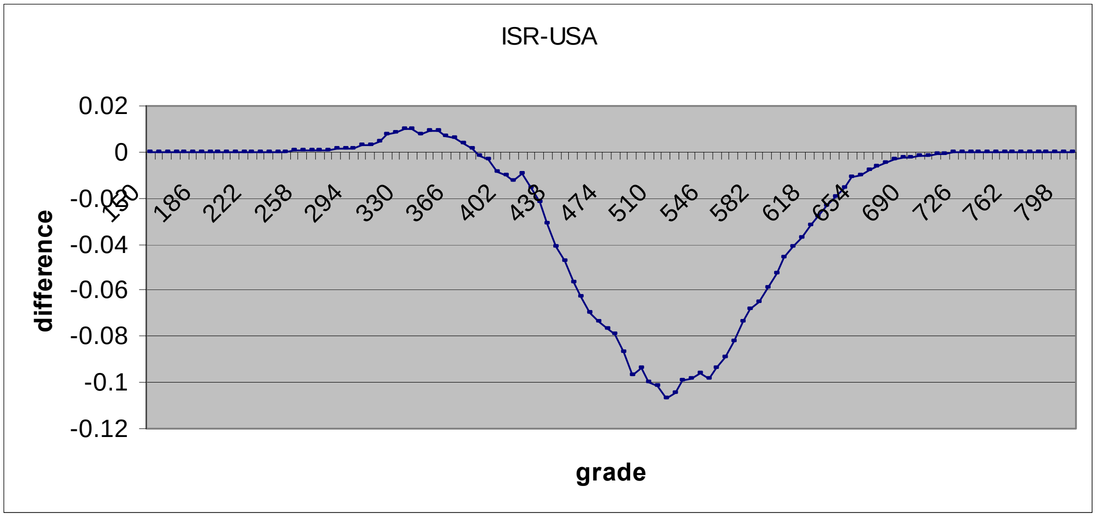
Table 2 is identical to Table 1 in structure.

Table 2. Descriptive statistics of grades of Israel and USA in Geometry

Group	Israel	USA
Sample size	4318	8912
Average	486.91	472.01
Standard deviation	79.61	63.74
Maximum grade	740.71	704.74
Minimum grade	242.43	245.54

According to Table 2 Israel is ranked (based on averages) higher than USA**.** Unlike Table 1, the worst grade belongs to the country with the higher average therefore it is easy to form a test which is composed of one question, such that the only person who does not answer correctly is the person with the lowest grade. Clearly, such an exam is expected to change the ranking of average grades. To look for less extreme examples, Figure 2 illustrates the difference between the cumulative distribution functions.

Figure 2. The cumulative distribution function for Israel minus the function for USA in Geometry



Following the features detailed in the first example it can be seen from Figure 2 that USA has relatively more students with higher grades than Israel in the range between 280 to 400 (the curve is positive)[14], while Israel has relatively more students with higher grades in the range of 400 to 710. Looking at the slopes of the curve, we can see that USA has relatively more students in the range 400 to 550, while Israel has relatively more students in the range 550 to 710. To search for an exam that will change the ranking of average grades we need more questions in the ability range that is related to 400 to 550 and less questions that will distinguish between abilities in the range 550 to 710. However, it is clear from the figure that it is harder (although possible) to find an exam that will change the ranking of average grades than it was in the earlier example because the distance of the curve from the horizontal axis is smaller. Also, as can be seen from the graph, the range over which the USA distribution represents higher percentage of students with higher ability is relatively small so that it will be harder to find a test that will reverse the average grades than in the case of Australia vs. Israel. (This is also indicated by the differences in average grades.

[14] In this range $F_I(g) > F_U(g)$ so that $1- F_I(g) < 1-F_U(g)$, which is the proportion of students with higher grades.

However, the difference in average grades need not be related to the difficulty of finding an alternative test, although it may be correlated with it). The p-value of the KS test is <0.001 therefore the hypothesis that the cumulative distributions are identical is rejected. This means that the hypothesis that the deviations of the curve from the horizontal axis are due to random fluctuations (due to reliance on a sample) is rejected.

<u>Example C: Algebra - Romania vs. Australia</u>

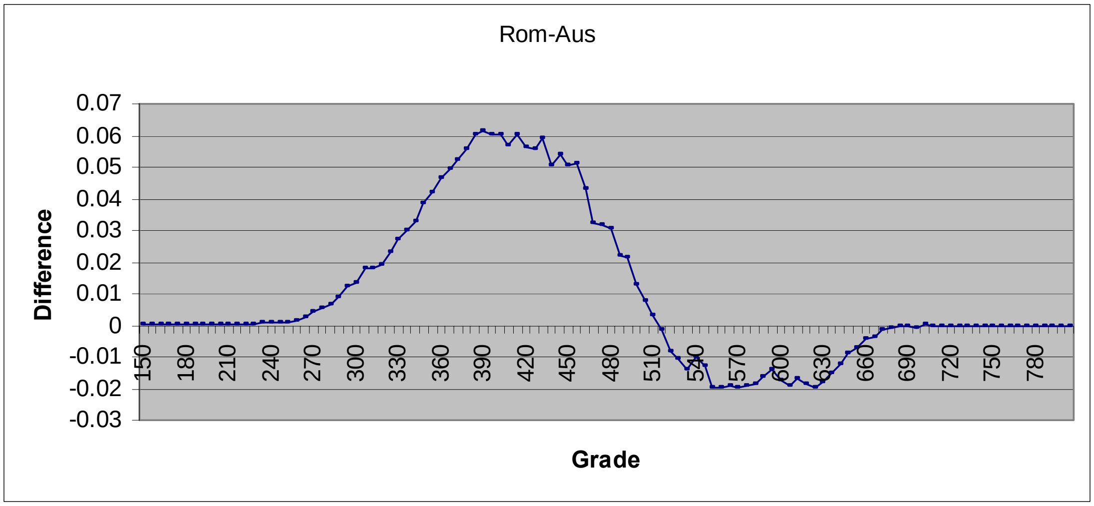
Table 3. Descriptive statistics of grades of Romania and Australia in Algebra

Group	Romania	Australia
Sample size	4104	4791
Average	484.88	491.43
Standard deviation	88.39	76.23
Maximum grade	748.17	745.08
Minimum grade	196.92	262.28

As can be concluded from Table 3, The quality of Australian students in algebra is higher on average than the quality of Romanian students. However, the best student is a Romanian. Therefore, one can design a questionnaire in which the Romanians will show a higher average grade. (A questionnaire that can be answered correctly only by the best Romanian student). Figure 3 presents the difference between the cumulative distributions.

It can be seen from Figure 3 that the range between 510 and 660 includes higher percentage of Romanian students with higher ability than the Australians, and therefore an exam that will include more difficult questions might improve the average scores of Romania relative to Australia, so that the ranking of average scores can change. Because the range of 510-660 is relatively large, one can expect that finding such tests is relatively simple. The p-value of the KS test is <0.001 therefore the hypothesis that the cumulative distributions are identical is rejected. This means that the hypothesis that the deviations of the curve from the horizontal axis are due to random fluctuations (due to reliance on a sample) is rejected.

Figure 3. The vertical difference between the cumulative distributions for Romania and Australia in Algebra



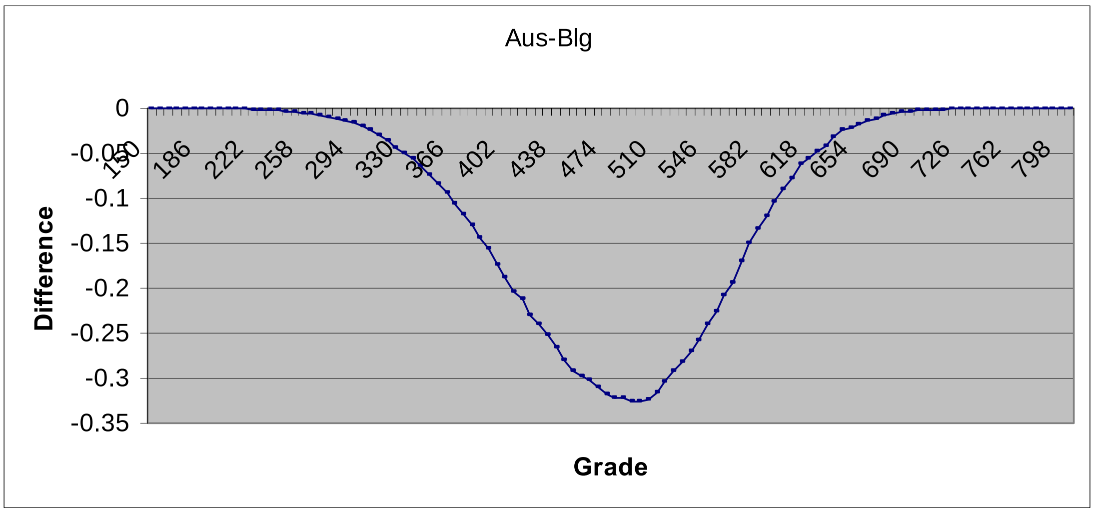
__**Example D: Data - Australia vs. Bulgaria**__

Table 4. Descriptive statistics of grades of Australia and Bulgaria in Data

Group	Australia	Bulgaria
Sample size	4791	4117
Average	527.21	465.39
Standard deviation	69.21	82.89
Maximum grade	755.95	736.20
Minimum grade	262.85	196.42

Example D is presented in order to indicate that a conclusion that one can **always** find an alternative test that can change the ranking of average grades is not correct. The Bulgarian empirical cumulative distribution is always higher than the Australian one (see Figure 4 below), making it impossible to find an alternative test that can change the ranking of average grades. One can argue that the large difference in average grades indicates that, but it is important to stress that it is not a sufficient condition for the ability to change the order of the average grades by an alternative test. The summary table in the Appendix contains cases with a difference in means as large as 36, where the condition for reversing the ranking is met (Israel vs USA in Data). Only the graph, presumably representing the ability distribution in the populations can answer such a question. The p-value of the KS test is <0.001 therefore the hypothesis that the cumulative distributions are identical is rejected.

Figure 4. The cumulative distribution function for Australia minus the function for Bulgaria in Data



Next, we summarize the results of all the comparisons performed in this paper. For the purpose of illustration we compared the performances of 5 countries (Australia, Bulgaria, Israel, Romania, and USA) in all the 6 types of exams.[15] Among the 60 comparisons, intersections were observed in 41 comparisons. The following table summarizes the results for the 5 countries (Australia, Bulgaria, Israel, Romania, and USA):

[15] This yields 60 comparisons $60 = 6 * \binom{5}{2}$ =6*10

Table 5. A summary of results for the 60 comparisons.

	KS significant	KS not significant	Total
Intersection found	39	2	41
Intersection not found	19	0	19
Total	58	2	60

Summarizing the table above we see that in 41 cases one could write a different exam, with a different difficulty distribution (sometimes easier, sometimes harder) and by doing that the order of ranking of average grades will be reversed. In 19 out of the comparisons there was no intersection. The detailed list of the results of the comparisons is presented in Appendix B. It can be seen that the ability to reverse the ranking of average grades, although correlated with, is not a simple function of the difference in average grades but it is related to the structure of the distributions and the way the difference in average grades is composed.

5. Conclusions and Suggestions for Further Research.

This paper points to a major defect in comparing average performance of groups in terms of a latent variable — e.g., ability. Even if the test and its procedure are agreed upon, changing the difficulty distribution of the questions in the questionnaire may cause reversal of the measured ability of groups, as measured by the mean scores of groups. It turns out that the conditions that enable mean reversal by changing the difficulty distribution are the same conditions that enable mean reversal by monotonic transformations of the latent variable. The paper offers a few examples that can indicate the probability of such an event occurring.

We used results from TIMSS to illustrate our point. For the 41 cases in which one can find an alternative test that can reverse the ranking of average grades it is clear that without further information, it is risky to reach definite conclusions with respect to the question which country is a better one. Our point in this paper is that the results of TIMSS, as all results that are based on average grades of a latent ability, should not be viewed as a result that came from a photo-finish analysis, and an analysis of the cumulative distributions should be carried out. This kind of analysis is illustrated in this paper. However, it is important to stress that in this paper we only looked at the **existence** of the possibility to change the ranking of the mean grades of groups, without being concerned with how hard it is to do so. Our guess is that the difficulty of finding an alternative test that can change the ranking is a function of ranges over which one distribution is below the other and the magnitude of this difference. Further research is needed to evaluate the probability of success, that is, how hard one has to search in order to find such an alternative exam.

It is worth pointing out that the possibility of mean reversal can spill over to other statistical methods. For example, consider an investigator who uses regression methods to estimate the production function of schooling. (To name a few – Kreuger (1999), Hanushek (1986), and Hanushek, Rivkin, and Taylor (1976). Under the circumstances described in this paper, it will be possible to reverse the sign of the regression coefficient by changing the difficulty distribution of the questionnaire (See Maddala, 1977, p. 162; Yitzhaki, 1990, and Yitzhaki and Schechtman (2004)). This means that researchers should (a) refrain from using grouped data, a point stressed in this paper or (b) check for the possibility of mean reversal before applying regression techniques. In other words, the same kind of reasoning that led to the results of this paper that aggregation may bias the results, aggregation in the form of regression can also lead to the possibility to change the sign of a regression coefficient between grades in a test and another variable of interest, like income. As far as we see, this is an important property to consider whenever there is an intention to examine the efficiency of different programs intended to improve grades or to evaluate different methods of teaching. Further research along the lines suggested in Yitzhaki (1990) and Yitzhaki and Schechtman (2004) is needed to formulate and examine this possibility.

To check the robustness of intersecting cumulative distributions, a test is needed. The KS test is not an adequate test for this purpose. A promising direction is to follow Schechtman et. al. (2008) which deals with tests for the intersection of Absolute Concentration Curves.

An additional and unrelated issue arises from our constraint on ability to be one-dimensional. In general we should expect multi-dimensional ability. Multi-dimensional ability can cause additional biases in group comparisons, like the Simpson paradox. Further research is needed in order to see whether the approach suggested in this paper can be of help in this area too. Assuming unidimensional ability implies that there should be a given structure among the responses to the questions intended to examine uni-dimensional ability. Ignoring random errors, the easiest question should be correctly answered by most participants, the second in the ranking of difficulty should be answered by a subset of those who answered the easiest question etc.. That is, if we find a question that is correctly answered by the worst and best students, while the middle ability group has failed, this is an indication that we have failed to identify a uni-dimensional ability.

A thorough empirical application of the propositions offered in the present study would inflate the paper beyond reason, and must be deferred to a later stage. Suitable databases exist, and statistical tests can be developed. As stressed in the introduction, the purpose of any such empirical application should be to uncover possible pitfalls inherent in the use of average scores in comparing groups, which result from tests' different difficulty distributions. Greater awareness of these hazards may contribute to greater efficiency in arriving at (budgetary) decisions that rely on such group comparisons.

References:

Barro, R. J. (1997). *Determinants of economic growth: A cross-country empirical study*. Cambridge, MA: MIT Press.

Barro, R. J. (2001). Human Capital and Growth, *American Economic Review*, 91, 2, (May), 12-17.

Brogden, H. E. (1977). The Rasch Model, The Law of Comparative Judgement and Additive Cojoint Measurement, *Psychometrika*, 42, 4 (December): 631–634.

Copeland, Thomas E. and J. Fred Weston (1983). *Financial Theory and Corporate Policy*, 2nd ed. Reading, MA: Addison-Wesley Publishing Company.

Ellis, Jules L. and Arnold L. van den Wollenberg (1993). Local Homogeneity in Latent Trait Models: A Characterization of the Homogeneous Monotone IRT Model, *Psychometrika*, 58 (No. 3, September): 417–429.

Kreuger, Alan B. (1999). Experimental Estimates of Education Production Functions, *Quarterly Journal of Economics*, 114, 2, 457, (May), 497-532.

Hanushek, Eric A. (1986). The Economics of Schooling: Production and Efficiency in the Public Schools, *Journal of Economic Literature*, 24 (No. 3, September): 1141–77.

——, Steven G. Rivkin, and Lori L. Taylor (1976). Aggregation and the Estimated Effects of School Resources, *The Review of Economics and Statistics*, 78, 4 (November): 611–627.

Huang, Chi-fu and Robert H. Litzenberger (1988). *Foundations for Financial Economics*. New York: North-Holland.

Levy, H. (2006). *Stochastic Dominance (investment decision making under uncertainty*. Second Edition, Springer.

Lehmann, E. L. (1955). Ordered Families of Distributions, *Annals of Mathematical Statistics*, 26: 399–419.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Maddala, G. S. (1977). *Econometrics*, New York, NY: McGraw-Hill Company.

Rasch, G. (1966). An Individualistic Approach to Item Analysis. In P. F. Lazarsfeld and N. W. Henry (eds.), *Readings in Mathematical Social Science*, Chicago: Science Research Associates.

Rubin, D. B., E. A. Stuart and E. L. Zanutto (2004). A potential outcomes view of value-added assessment in education, *Journal of Educational and Behavioral Statistics,* 29, 103-116.

Saposnik, R. (1981). Rank Dominance in Income Distribution, *Public Choice*, 36: 147–151.

Schechtman, E. and S. Yitzhaki (2006). Ranking groups' abilities – is it always reliable? Mimeo http://papers.ssrn.com/sol3/papers.cfm?abstract_id=938529

Schechtman, E.; A. Shelef; S. Yitzhaki and R. Zitikis (2008). Testing hypothesis about absolute concentration curve and marginal conditional stochastic dominance, *Econometric Theory*, 24,. 4. Forthcoming.

Spencer, Bruce D. (1983a). On Interpreting Test Scores as Social Indicators: Statistical Considerations, *Journal of Educational Measurement*, 20, 4 (Winter): 317–333.

––– (1983b). Test Scores as Social Statistics: Comparing Distributions, *Journal of Educational Statistics*, 8: 249–269.

Terwilliger, J. and M. Schield (2004). Frequency of Simpson's Paradox in NAEP Data, AERA 2004, 4/9/2004, mimeo.

Wainer, H. (1986a). Minority contributions to the SAT score turnaround: an example of Simpson's Paradox, *Journal of Educational Statistics*, 11, 229-244.

Wainer, H. (1986b). The SAT as a social indicator: a pretty bad idea, in Wainer, H. (ed.) *Drawing Inference from Self-Selected Samples*, Springer-Verlag:

Wainer, H. (1994). On the academic performance of New Jersey's public school children: fourth and eighth grade Mathematics in 1992, *Education Policy Analysis Archives*, 2, 10, http://epaa.asu.edu/epaa/v2n10.html

Wainer, H. and L. M. Brown (2004). Two statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data, *The American Statistician*, 58, 2, (May), 117-123.

Wainer, H. & Brown, L. (2007). Three statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. Ch. 26, pp 893-918, in *Handbook of Statistics* (27), *Psychometrics* (Eds. C. R. Rao and S. Sinharay). Elsevier Science: Amsterdam.

Yitzhaki, S. (1990). On the sensitivity of a regression coefficient to monotonic transformations, *Econometric Theory*, 6, No. 2, 165-169.

Yitzhaki, S. and M. Eisenstaedt (2003). Groups' versus individuals' ranking. In Amiel, Yoram and John A. Bishop (eds.) *Fiscal Policy, Inequality, and Welfare, Research on Economic Inequality*, 10, Amsterdam: JAI, 101-123.

Yitzhaki, S. and E. Schechtman (2004). The Gini Instrumental Variable, or the "double instrumental variable" estimator, *Metron*,, LXII, 3, 287-313.

Appendix A:

The mathematics assessment framework for TIMSS 2003 is framed by two organizing dimensions, a content dimension and a cognitive dimension.

The mathematics content domains:

Number

The number content domain includes understanding of counting and numbers, ways of representing numbers, relationships among numbers, and number systems.

The number content domain consists of understandings and skills related to:

1. whole numbers
2. fractions and decimals
3. integers
4. ratio, proportion, and percent

Algebra

The algebra content domain includes patterns and relationships among quantities, using algebraic symbols to represent mathematical situations, and developing fluency in producing equivalent expressions and solving linear equations.

This domain include:

1. patterns
2. algebraic expressions
3. equations and formulas
4. relationships

Measurement

Measurement involves assigning a numerical value to an attribute of an object. The focus of this content domain is on understanding measurable attributes and demonstrating familiarity with the units and processes used in measuring various attributes.

The measurement content domain is comprised of the following two main topic areas:

1. attributes and units
2. tools, techniques, and formula

Geometry

The geometry content area includes understanding coordinate representations and using spatial visualization skills to move between two- and three-dimensional shapes and their representations. Students should be able to use symmetry and apply transformation to analyze mathematical situations.

The major topic areas in geometry are:

1. lines and angles
2. two- and three-dimensional shapes
3. congruence and similarity
4. locations and spatial relationships
5. symmetry and transformations.

Data

The data content domain includes understanding how to collect data, organize data that have been collected by oneself or others, and display data in graphs and charts that will be useful in answering questions that prompted the data collection. This content domain includes understanding issues related to misinterpretation of data (e.g., about recycling, conservation, or manufacturers' claims).

The data content domain consists of the following four major topic areas:

1. data collection and organization
2. data representation
3. data interpretation
4. uncertainty and probability.

The mathematics cognitive domains:

Knowing Facts and Procedures

Facts encompass the factual knowledge that provides the basic language of mathematics, and the essential mathematical facts and properties that form the foundation for mathematical thought.

Procedures form a bridge between more basic knowledge and the use of mathematics for solving routine problems, especially those encountered by many people in their daily lives.

Using Concepts

Familiarity with mathematical *concepts* is essential for the effective use of mathematics for problem solving, for reasoning, and thus for developing mathematical understanding.

Knowledge of concepts enables students to make connections between elements of knowledge that, at best, would otherwise be retained

as isolated facts. It allows them to make extensions beyond their existing knowledge, judge the validity of mathematical statements and methods, and create mathematical representations.

Representation of ideas forms the core of mathematical thinking and communication, and the ability to create equivalent representations is fundamental to success in the subject.

Solving Routine Problems

The routine problems will have been standard in classroom exercises designed to provide practice in particular methods or techniques. Some of these problems will have been in words that set the problem situation in a quasi-real context. Solution of other such "textbook" type problems will involve extended knowledge of mathematical properties (e.g., solving equations). Though they range in difficulty, each of these types of "textbook" problems is expected to be sufficiently familiar to students that they will essentially involve selecting and applying learned procedures.

Reasoning

Reasoning mathematically involves the capacity for logical, systematic thinking. It includes intuitive and inductive reasoning based on patterns and regularities that can be used to arrive at solutions to non-routine problems. Non-routine problems are problems that are very likely to be unfamiliar to students. They make cognitive demands over and above those needed for solution of routine problems, even when the knowledge and skills required for their solution have been learned.

The data contained in the site which we analyze include students' performance in the content domain and also in mathematics overall.

Appendix B:

The list of all comparisons performed is presented in the following Table:

Results of Comparisons According to Exam and Country

Domain	Countries	No. of observations	Average grades	Possible to change?	p-value for KS Statistic
Math overall	(Romania, Australia)	(4104,4791)	(480.1,497.6)	no	<0.001
Math overall	(Romania, Bulgaria)	(4104,4117)	(480.1,483.5)	yes	0.015
Math overall	(Romania, Israel)	(4104,4318)	(480.1,495.1)	no	<0.001
Math overall	(Australia, Bulgaria)	(4791,4117)	(497.6,483.5)	yes	<0.001
Math overall	(Australia, Israel)	(4791,4318)	(497.6,495.1)	yes	0.022
Math overall	(Bulgaria, Israel)	(4117,4318)	(483.5,495.1)	yes	<0.001
Math overall	(Romania, USA)	(4104,8912)	(480.1,504.1)	no	<0.001
Math overall	(Australia, USA)	(4791,8912)	(497.6,504.1)	yes	<0.001
Math overall	(Bulgaria, USA)	(4117,8912)	(483.5,504.1)	yes	<0.001
Math overall	(Israel, USA)	(4318,8912)	(495.1,504.1)	yes	<0.001
Algebra	(Romania, Australia)	(4104,4791)	(484.9,491.4)	yes	<0.001
Algebra	(Romania, Bulgaria)	(4104,4117)	(484.9,486.3)	yes	<0.001
Algebra	(Romania, Israel)	(4104,4318)	(484.9,497.1)	yes	<0.001
Algebra	(Australia, Bulgaria)	(4791,4117)	(491.4,486.3)	yes	<0.001
Algebra	(Australia, Israel)	(4791,4318)	(491.4,497.1)	yes	<0.001
Algebra	(Bulgaria, Israel)	(4117,4318)	(486.3,497.1)	no	<0.001
Algebra	(Romania, USA)	(4104,8912)	(484.9,509.9)	yes	<0.001
Algebra	(Australia, USA)	(4791,8912)	(491.4,509.9)	yes	<0.001
Algebra	(Bulgaria, USA)	(4117,8912)	(486.3,509.9)	no	<0.001
Algebra	(Israel, USA)	(4318,8912)	(497.1,509.9)	yes	<0.001
Number	(Romania, Australia)	(4104,4791)	(479.1,490.9)	no	<0.001
Number	(Romania, Bulgaria)	(4104,4117)	(479.1,483.8)	yes	0.01
Number	(Romania, Israel)	(4104,4318)	(479.1,503.0)	no	<0.001
Number	(Australia, Bulgaria)	(4791,4117)	(490.9,483.8)	yes	<0.001

Domain	Countries	No. of observations	Average grades	Possible to change?	p-value for KS Statistic
Number	(Australia, Israel)	(4791,4318)	(490.9,503.0)	yes	<0.001
Number	(Bulgaria, Israel)	(4117,4318)	(483.8,503.0)	yes	<0.001
Number	(Romania, USA)	(4104,8912)	(479.1,507.3)	no	<0.001
Number	(Australia, USA)	(4791,8912)	(490.9,507.3)	yes	<0.001
Number	(Bulgaria, USA)	(4117,8912)	(483.8,507.3)	yes	<0.001
Number	(Israel, USA)	(4318,8912)	(503.0,507.3)	yes	0.0547
Geometry	(Romania, Australia)	(4104,4791)	(480.0,482.9)	yes	<0.001
Geometry	(Romania, Bulgaria)	(4104,4117)	(480.0, 491.4)	no	<0.001
Geometry	(Romania, Israel)	(4104,4318)	(480.0,486.9)	no	<0.001
Geometry	(Australia, Bulgaria)	(4791,4117)	(482.9, 491.4)	yes	<0.001
Geometry	(Australia, Israel)	(4791,4318)	(482.9,486.9)	yes	<0.001
Geometry	(Bulgaria, Israel)	(4117,4318)	(491.4, 486.9)	yes	0.0718
Geometry	(Romania, USA)	(4104,8912)	(480.0,472.0)	yes	<0.001
Geometry	(Australia, USA)	(4791,8912)	(482.9,472.0)	yes	<0.001
Geometry	(Bulgaria, USA)	(4117,8912)	(491.4,472.0)	yes	<0.001
Geometry	(Israel, USA)	(4318,8912)	(486.9,472.0)	yes	<0.001
Data	(Romania, Australia)	(4104,4791)	(450.4,527.2)	no	<0.001
Data	(Romania, Bulgaria)	(4104,4117)	(450.4,465.4)	no	<0.001
Data	(Romania, Israel)	(4104,4318)	(450.4,490.0)	no	<0.001
Data	(Australia, Bulgaria)	(4791,4117)	(527.2,465.4)	no	<0.001
Data	(Australia, Israel)	(4791,4318)	(527.2,490.0)	yes	<0.001
Data	(Bulgaria, Israel)	(4117,4318)	(465.4,490.0)	no	<0.001
Data	(Romania, USA)	(4104,8912)	(450.4,526.4)	no	<0.001
Data	(Australia, USA)	(4791,8912)	(527.2,526.4)	yes	0.0389
Data	(Bulgaria, USA)	(4117,8912)	(465.4,526.4)	no	<0.001
Data	(Israel, USA)	(4318,8912)	(490.0,526.4)	yes	<0.001
Measurement	(Romania, Australia)	(4104,4791)	(488.8,503.6)	no	<0.001

Domain	Countries	No. of observations	Average grades	Possible to change?	p-value for KS Statistic
Measurement	(Romania, Bulgaria)	(4104,4117)	(488.8,478.1)	yes	<0.001
Measurement	(Romania, Israel)	(4104,4318)	(488.8,480.5)	yes	<0.001
Measurement	(Australia, Bulgaria)	(4791,4117)	(503.6,478.1)	yes	<0.001
Measurement	(Australia, Israel)	(4791,4318)	(503.6,480.5)	no	<0.001
Measurement	(Bulgaria, Israel)	(4117,4318)	(478.1,480.5)	yes	<0.001
Measurement	(Romania, USA)	(4104,8912)	(488.8,495.0)	yes	<0.001
Measurement	(Australia, USA)	(4791,8912)	(503.6,495.0)	yes	<0.001
Measurement	(Bulgaria, USA)	(4117,8912)	(478.1,495.0)	yes	<0.001
Measurement	(Israel, USA)	(4318,8912)	(480.5,495.0)	yes	<0.001

It can be seen from the Table that even if the difference in average grades is relatively large it may still be possible to find an alternative test that will change the ranking of average grades. For example, although the difference between Australia and Israel in Data is relatively large (527 vs. 490) it is still possible to reverse the ranking of average grades. On the other hand, although the difference between Romania and Australia (479 and 490 respectively) in Number is relatively small, it is impossible to find an alternative test that will reverse the ranking.

תקציר

השוואות של ממוצעי הציונים שהשיגו תלמידים בתחומים של מתמטיקה מהוות היום בסיס להשוואה בין מדינות ולמקור גאווה או זעקות על כשלון מערכתי של מערכות חינוך ותופסות כותרות בעמודים הראשיים של עיתונים. מטרתם של מבחנים אלו היא לאתר את הידע "הממוצע" שיש לנבחנים. ידע, בשונה ממשתנה מדיד כגון גובה, אינו נצפה ישירות ועל המודד לנסח מבחן עם שאלות בדרגות קושי שונות שמאפשר לראות את המידה בה ידעו הנבחנים לענות על השאלות. הבדל זה בין מדידת גובה ומדידת ידע גורר שבעוד שהחוקרים מתייחסים לשני התחומים בצורה זהה הרי שהידע הממוצע, בניגוד לגובה ממוצע, תלוי במבנה השאלון ובהתפלגות הקושי של השאלות בשאלון. בעבודה זו אנו מציעים שיטה לבחון את אפשרות הקיום של מבחן אחר, שגם הוא מודד ידע, ושתוצאתו תגרום להיפוך של הסדר של ממוצעי הציונים בין התלמידים במדינות השונות. כלומר, המדינה שקיבלה במבחן הרשמי את הציון הממוצע הגבוה יותר תקבל במבחן האלטרנטיבי ציון שאינו הגבוה ביותר.

השיטה המוצעת הודגמה על נתוני מבחן בינלאומי שנקרא TIMSS. בעבודה זו השווינו את תוצאות המבחנים של תלמידים מחמש מדינות: אוסטרליה, ארצות הברית, בולגריה, ישראל ורומניה, כאשר בכל מדינה קיימות תוצאות של שישה מבחנים בתחומי המתמטיקה השונים. מתוך 60 ההשוואות האפשריות מצאנו שבארבעים ואחד מקרים, ניתן היה למצוא מבחן אחר שהיה משנה את הסדר של המדינות. המסקנה המתבקשת היא שבהשוואת ממוצעי ציונים יש לבדוק את מידת העמידות של הממצאים בהתאם לשיטה המוצעת במאמר.

<u>מילות מפתח:</u> יכולת, מדידה, TIMSS

הוצאת הלשכה המרכזית לסטטיסטיקה, רח' כנפי נשרים 66, פינת רח' בקי,
ת"ד 34525, ירושלים 91342
טל': 02-6592666; פקס: 02-6521340
אתר הלמ"ס באינטרנט: www.cbs.gov.il
דואר אלקטרוני: info@cbs.gov.il

WORKING PAPER SERIES

NO. 37

The Robustness of Conclusions Based on TIMSS Mean Grades

Edna Schechtman, Efrat Soffer** and Shlomo Yitzhaki***

July 2008

*Central Bureau of Statistics – Director of the Bureau
**Ben-Gurion University