Fifteen-Years-Old Students of Seven East Asian Cities in PISA 2009: A Secondary Analysis

SOH Kay Cheng
Independent Consultant, Singapore

Abstract

Background: In PISA 2009, seven East Asian countries rank high among the 65 participating countries, but some of the differences among the seven countries are small to be of substantive meaning.

Aims: This paper is an attempt to fine tune the comparisons for better understanding of the situation in East Asian.

Sample: Data of the seven East Asian countries were pulled from the PISA 2009 report and re-analyzed.

Method: Pair-wise comparisons were made by way of effect size on Reading, Mathematics, and Science.

Results: The overall patterns of differences show that Shanghai-China is definitely ahead of all the others. Korea, Hong Kong-China, and Singapore are similar in performance and form a cluster. Japan, Chinese Taipei, and Macao-China are similar and form another cluster at the lower end of achievement.

Conclusion: Instead of ranking the seven countries with seven different ranks, it is more meaningful to cluster them into three groups to avoid spurious precision. In other words, league tables should not differentiate where there are no meaningful differences.

Keywords: International comparisons; East Asian; effect size

東亞七國十五歲學生在PISA表現的再度分析

蘇啟禎
獨立顧問，新加坡

摘要

背景：PISA 2009的調查結果, 東亞七國在65參與國當中列位很高。但其中有些差異太小而毫無實質意義。

目的：本文嘗試比較細緻的分析，以期促進對東亞各國的情況有更確切的瞭解。

研究物件：PISA 2009報告書中有關東亞七國的資料被採用，作再度分析。

研究方法：針對閱讀、數學、與科學三科，採用效果強度進行配對比較。

結果：整體的差異傾向顯示，上海遙遙領先，而韓國、香港、和新加坡表現接近，成為一組。日本、臺北、和澳門表現也相近，成為較低的一組。

結論：東亞七國不應該有七個列位，而歸納為三組則比較合理，以避免虛無的準確性。

關鍵詞：國際比較、東亞、效果強度

Introduction

In the *Foreword* of the *PISA 2009 Results: What Students Know and Can Do -- Student performance in Reading, Mathematics, and Science, Volume 1*, the OECD Secretary-General, Angel Gurría makes it abundantly clear that the Programme for *International Student Assessment* (PISA) is to provide "*reliable information on how well education systems prepare students for life*." Thus, PISA evaluates the quality, equity and efficiency of school systems in 65 countries/cities (for brevity, hereafter *cities*).

As in other international studies of similar nature (e.g., TIMSS, IEA 2011), PISA 2009 reports that East Asian cities score at a high level. And, as is true of many international comparisons (e.g., university ranking, competitiveness, corruption, etc.), the outcomes become a league tables, although this might not be the original intention of the studies (Soh, 2011a). Thus, the focus is on *Have we done better than others* and *Why we are doing well (or not so well) what we are doing*. Then, educational research becomes an international contest like the Olympic and the Miss World pageant, although the original intent is to enable participating cities to evaluate their educational achievement. This competitive approach seems to go against the spirit of PISA as cited above.

When cities focus on positions in educational league table, the conceptual and technical aspects are usually neglected or ignored, and the outcomes are taken literally with no safeguard against possible misconstrue. On the other hand, when taken with due cautions to its limitations inherent in data analysis and presentation, the reported outcomes are still able to inform policy-making (Mortimore, 2009).

As mentioned, the seven East Asian cities have all done well in PISA 2009, ranked high in the international league table. There is however a value in comparing them within the regional context. The objective of this secondary analysis is, therefore, to facilitate this comparison to see if relevant differences do exist. Being a secondary analysis, this paper by definition does not present new data but a more circumscribed perspective. It is hoped that new insights will be brought about for East Asian educationists to better understand their achievement which has impressed the world over.

Methodological Issues

Before presenting the results of the secondary analysis, a few methodological issues need be highlighted as a caution against unreserved confidence.

Language of Testing

PISA assessed the 15-year-olds in Reading with 131 items, Mathematics with 35 items, and Sciences with 53 items. The items have a variety of format, content, and context to take into consideration the complexity of learning experiences and outcomes. A particular interest and strength of it is the emphasis on process skills in Reading supposedly demanded by the world of work in the 21st Century.

The three subject tests were administered in the students' language of instruction, which in most cases is also the home language (or mother tongue). It is therefore safe to assume that the Japan sample was tested in Japanese and Korea sample in Korean, etc., not only for Reading but also Mathematics and Science. However, this may not be the case for Singapore where complex language background of the students in a classroom is the rule than exception and there is a sizeable proportion of students whose home language is *not* the language of instruction, English in this case.

It is readily appreciated that proficiency in the language of instruction can impact on the test

performance and in the Singapore's case student performance might have been thus curtailed to an unknown extent. And, the extent to which this is true is a topic for further investigation. More specifically, according to the 2010 census (Singapore Department of Statistics, 2010), among Singapore residents aged 15 to 24, 41% of Chinese spoke English most frequently as home language. The corresponding figures for Malay are 18% and Indian 52%. Note that the proportions of the three ethnic groups are 74%, 13%, and 9%.

Significance Testing

The PISA reports results of null hypothesis significance test (NHST) for comparing cities as being 'significantly different' or otherwise. The NHST has a history of more than 80 years and its usefulness has been a controversy for the past decades on which a tremendous amount of articles and monographs have been published (e.g., Harlow, Mulaik, & Steiger, 1997). The NHST (the t-test being its archetype) answers the question of the *probability of chance occurrence of an obtained difference* when one is obtained and not answering the question of *the magnitude of the obtained difference*. Abelson (1995, p. 40) illustrates the function of the NHST (e.g., the ubiquitous t-test) by giving an example of an obtained differences which is said to be '*significant at the 0.01 level'* thus,

> If it were true that there were no systematic difference between the means in the population from which the samples came, then the probability that the observed means would have been as different as they were, or more different, is less than one in a hundred. This being strong grounds for doubting the viability of the null hypothesis, the null hypothesis is rejected.

From the educational practitioner's viewpoint, the concern is the *magnitude* of an observed difference and not the *probability* of it. The prevalent and persistent interest in NHST might have been generated, at least partly, by the confusion of *statistical* meaning with the *lay* meaning of the word *significance* and its derivative *significant.* For instance, two very large groups of students are found to have a small *significant* difference and it is taken to be *important* because that small difference is reported as being *significant*. There is obviously a confusion of statistical significance with educational significance (Soh, 2011b).

Moreover, statistical significance is a function of sample size, *inter alia*. In other words, the statistical significance of t-value is confounded by sample size. A small difference between two small groups is statistically non-significant and hence is likely ignored. However, the *same* small difference is 'statistically significant' when the sample sizes are large enough and is likely to be taken seriously as indicating a truly important difference. Thus, the confusion of technical and daily uses of the terms *significance* and *significant* leads to the confusion between statistical and practical importance of an obtained difference, irrespective of its being large, medium, or small in magnitude.

Effect Size

It is with such concern that the American Psychological Association adopted the recommendation and then recommends in its fifth *Publication Manual* that research be reported with effect size, in addition to the traditional HNST results. Since then, more than 25 important learned journals in education, psychology and related fields have made this their publication policy (Thompson, 1998). However, the PISA report does seem to have taken this into consideration and there is hardly,

if any, mention of effect size. The *magnitude* indicated by the effect size for a between-nation difference should be of greater concern to educational practitioners and policy-makers than the *probability* indicated by the t-value and its corresponding p-value. For this reason, in this secondary analysis, the seven East Asian cities are compared by way of effect size.

There are several formulae of effect size in terms of standardized mean difference (SMD) for different theoretical concerns and purposes. They however, yield effect sizes which differ in the second and even the third decimal values. Thus, for practical purposes, they can be taken to be equivalents (Soh, 2008). The formula used in this secondary analysis is Glass's *delta* which is the most straightforward as shown below:

Effect size (SMD, Glass's *delta*) = $(Mean_1 - Mean_2)/SD_2$

Effect sizes are evaluated for their magnitudes using Cohen's (1988) criteria below:

0.8 and above	Large effect
0.5 to 0.8	Medium effect
0.2 to 0.5	Small effect
0.0 to 0.2	Trivial effect

Data Analysis and Results

Reading

Table 1 shows the performance levels and effect sizes for Reading among the seven East Asian cities and effect sizes equal to or greater than 0.80 are highlighted. As can be seen therein, Shanghai-China topped the list, scoring higher than five other East Asian cities, except Korea. Korea. At the same time, Hong Kong-China scored higher than did Japan, Chinese-Taipei, and Macao-China. And, Singapore and Japan both out-scored Chinese-Taipei and Macao-China, while the last two cities did not differ.

Table1

Reading Performance and Effect Sizes

	Mean (Rank)	**SD**	**Effect Size**						
			Shanghai-China	Korea	Hong Kong-China	Singapore	Japan	Chinese Taipei	Macao-China
Shanghai-China	550 (1)	17	0.00						
Korea	539 (2)	25	0.64	0.00					
Hong Kong-China	533 (3)	14	**0.99**	0.24	0.00				
Singapore	526 (4)	8	**1.40**	0.53	0.49	0.00			
Japan	520 (5)	27	**1.75**	0.77	**0.92**	0.75	0.00		
Chinese Taipei	495 (6)	20	**3.20**	**1.78**	**2.69**	**3.88**	**0.92**	0.00	
Macao-China	487 (7)	7	**3.67**	**2.10**	**3.25**	**4.88**	**1.21**	0.40	0.00

Mathematics

Table 2 shows the performance levels and effect sizes of the differences in Mathematics, with effect sizes indicating large (0.80) or near large (0.75) differences highlighted. For this subject, Shanghai-China outperformed all six other East Asian cities.

Korea performed on par with the other five East Asian cities. However, Hong Kong-China scored higher than did Japan and Macao-China, while Singapore out-performed Japan, Chinese-Taipei, and Macao-China. The last three cities did not differ in the performance of Mathematics.

Table 2

Mathematics Performance and Effect Sizes

	Mean (Rank)	**SD**	**Effect Size**						
			Shanghai-China	Korea	Hong Kong-China	Singapore	Japan	Chinese Taipei	Macao-China
Shanghai-China	600 (1)	20	0.00						
Korea	546 (4)	28	**2.70**	0.00					
Hong Kong-China	555 (3)	18	**2.25**	-0.32	0.00				
Singapore	562 (2)	10	**1.90**	-0.57	-0.38	0.00			
Japan	529 (6)	26	**3.55**	0.60	**1.43**	**3.24**	0.00		
Chinese Taipei	543 (5)	26	**2.85**	0.11	0.66	**1.87**	-0.54	0.00	
Macao-China	525 (7)	7	**3.75**	0.74	**1.65**	**3.64**	0.16	0.69	0.00

Science

Table 3 shows the performance levels and effect sizes of the differences in Science. Shanghai-China again topped the list and is followed by Hong Kong-China and Singapore which, in turn, scored higher than Japan and Macao-China. Korea scored higher than did Macao-China, and Chinese Taipei outperformed Macao-China. At the same time, Japan, Chinese Taipei, and Macao-China performed equally well.

Table 3

Science Performance and Effect Sizes

	Mean (Rank)	SD	Effect Size						
			Shanghai-China	Korea	Hong Kong-China	Singapore	Japan	Chinese Taipei	Macao-China
Shanghai-China	575 (1)	16	0.00						
Korea	538 (5)	24	**2.25**	0.00					
Hong Kong-China	549 (2)	19	**1.58**	-0.46	0.00				
Singapore	542 (3)	10	**2.01**	-0.17	0.37	0.00			
Japan	530 (6)	27	**2.74**	0.33	**1.01**	**1.18**	0.00		
Chinese Taipei	539 (4)	20	**2.19**	-0.04	0.53	0.29	-0.34	0.00	
Macao-China	515 (7)	8	**3.65**	**0.96**	**1.80**	**2.65**	0.57	**1.21**	0.00

Overall

Table 4 shows the overall performance and effect sizes when the means (and SDs) for the three subjects are combined. As would therefore be expected, Shanghai-China maintained the top position among the seven East Asian cities on this measure, followed by Hong Kong-China, Singapore, and Korea. This is followed by Japan and Chinese Taipei which shared the same position of 5.5, and the list ends with Macao-China. Avoiding spurious precision, the natural groupings of the cities are [Shanghai-Chinese], [Hong Kong-China, Singapore, Korea], [Japan, Chinese Taipei], and [Macao-China].

In term of differences in effect sizes, Shanghai-China out-scored all other six East Asian cities, while Hong Kong-China and Singapore did so with Japan, Chinese Taipei, and Macao-China. Korea outperformed only Macao-China, while Chinese Taipei did likewise with Macao-China. Japan, Chinese Taipei, and Macao-China performed on par with one another.

Table 4

Overall Performance and Effect Sizes

	Mean (Rank)	SD	Effect Size						
			Shanghai-China	Korea	Hong Kong-China	Singapore	Japan	Chinese Taipei	Macao-China
Shanghai-China	575 (1)	18	0.00						
Korea	541 (4)	26	**1.89**	0.00					
Hong Kong-China	546 (2)	17	**1.63**	-0.18	0.00				
Singapore	543 (3)	10	**1.76**	-0.09	0.14	0.00			
Japan	526(5.5)	27	**2.71**	0.57	**1.12**	**1.79**	0.00		
Chinese Taipei	526(5.5)	22	**2.75**	0.60	**1.16**	**1.86**	0.03	0.00	
Macao-China	509 (7)	7	**3.68**	**1.24**	**2.13**	**3.61**	0.65	0.77	0.00

Discussion and Conclusion

PISA begins with the objective of providing reliability and valid information of student achievement to assist in educational decision-making. This seemingly simple task turns out to be rather complicated as attested by four volumes of PISA reports. The secondary analysis reported here is only one of many possible further working on the information thereof, similar to studies based information available from census reports. As hindsight always looks wiser, there are several points arising from this analysis calling for some discussion.

Spurious Precision and Ranking

Three points need be mentioned here. First, any two scores with a difference can be ranked to index the difference although it may be a minute one with little or no substantive meaning. For instance, in overall performance, Hong Kong-China is ranked second, Singapore third, and Korea fourth, but their scores (respectively, 546, 543, and 541, as shown in the PISA Report) differ so little that the differences

have little substantive value and hence are best ignored.

Secondly, the same rank differences may have different meanings. For instance, based on overall means, Hong Kong-China, Singapore, and Korea are ranked second, third, and fourth with their overall means being 546, 543, and 541, respectively. The differences are 3 and 2. In both cases, the ranks differences are one. The problem of ranking based on spurious precision is clearly demonstrated here when the unit of measure are not explicitly stated. When it is reported that two entities differ by '1', very little attention will be paid to the difference. How, when the same two entities are reported to have a difference of, for example, '830', heads will raise. The problem here is that one SGD is worth 830 Korean Won when units are specified. The importance of unit is obvious. But, for educational measures like the PISA means, the unit is not specified.

Thirdly, ranking is relative and not absolute; the best may not be good enough and the worst may not be so bad. For instance, Shanghai-China is the best on all four measures, but in view of the possible highest score of 800 on the long scale (mean 500 and SD 100 as used in PISA), the highest mean of 575 for Overall suggests that there is still much room for improvement; the best may not be good enough. On the other hand, Macao-China consistently comes last on the list of the seven East Asian cities. However, her overall mean of 509 places her exactly at the middle of the scale in the internal context of 65 cities; the worst may not be so bad. The limitation of ranking is obvious.

In short, when using ranking for comparison (in PISA and any other matter which matters), it is prudent to look not only at the relative positions but also ask what the ranks (and the differences on which ranking is based) really represent in substantive terms. Moreover, to avoid the pitfall of spurious precision, of seeing a difference when there is substantively none, grouping that ignores minute differences reflect the situation more accurately. Ironically, it takes less accuracy to be more accurate! The problem of spurious precision has recently become a concern in many fields other than education, such as geography (Foote & Huebner, 1995), health and science (Revere, 2011), and even law (Morrison, 2006). Education needs to catch up with these disciplines.

Nations or Cities

In PISA 2009, the People's Republic of China (PRC) is represented by three cities as independent entities. They consistently occupy the top, the middle, and the lowest positions in the achievement ranking among the seven Asian cities. Had Shanghai-China, Hong Kong-China, and Macao-China been combined to represent the PRC, the results will be quite different. On the other hand, Singapore is a city *and a nation at the same time* and, as can be expected as well as shown by PISA, there is a much narrower range in the talent pool. The question is, how meaningful can comparisons be made between these four cities and those that are truly nations (e.g., Japan and Korea). Of course, this depends on the definitions of *city* and *nation* and also on the sampling procedure; an issue which may need be sorted out in future international studies, be it PISA or others.

Language of Testing

PISA administered the tests in the students' medium of instruction. The tacit assumption is that the students are proficient in that language and their performance would not be influenced adversely. However, as alluded to earlier, diversity

in language background can be expected to impact on performance. A case in point is Singapore where language diversity is the rule than exception in that there is always a sizeable proportion of students whose home language is not also the medium of instruction used in the PISA assessment. This may and may not be unique issue to Singapore as bilingualism (or even multilingualism) is a norm of to-day's world.

Related to this is the equivalence of translated tests since some many cities with different languages are involved in PISA (and similar international studies). Van der Vijer & Hambleton (1996) differentiate between three distinct types of bias related to test translation that may affect the validity of tests adapted from different cultural contexts: construct, method, and item biases. It is easy to imagine the difficulties when translating the tests that will fit all the linguistic, cultural, and social contexts. To illustrate, Robinson (2010) found Spanish-speaking English Language Learners in kindergarten and first grade performed better on mathematics assessments when tested in Spanish, instead of English, where effect sizes were greater than 0.85. Admittedly, it is not an easy task for PISA to translate the three subject tests into so many languages to match the media of instruction of the participating cities. For their effort in this, the PISA is to be commended. Nonetheless, in the context of PISA, this is a topic worthy of further research effort.

Tremendous resources have been put into international studies like PISA. As long as there is an interest, such activities will continue, notwithstanding conceptual and technical issues to which rank-users (especially, educational administrators, politicians, and policy-makers) usually are oblivious or prefer to ignore. The outcomes are supposed to help participating cities (or nations) to evaluate students' achievement in the widest context on earth and to benefit from the evaluation. This, therefore, calls for not only active participation but also proper utilization of the information so costly obtained, definitely more than merely answering a question such as *how do we compare with others?*

Nevertheless, proper utilization of information needs be predicated by proper understanding and interpretation of the information made available by such studies of a mammoth scale. And, as this secondary analysis demonstrates, this is not as simple as subtracting one rank from another. It may therefore be apt to conclude by citing one of the most prolific and popular writers of science fiction of the 20th Century, P. Anderson:

> I have yet to see any problem, however complicated, which, when looked at in the right way did not become still more complicated.

References

Abelson, R. P. (1995). *Statistcs as Principled Argument*. Hillsdale, NJ: Erlbaum.

Cohen, J,. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition. Hillsdale, NJ: Erlbaum.

Harlow, L. L., Mulaik, S. A., & Steiger, L. H. (Eds). (1997). *What If There Were No Significance Tests?* Mahwah, NJ: Erlbaum.

International Association for the Evaluation of Educational Achievement, IEA (2011). *Trends in International Mathematics and Science Study 2011*. Retrieved on 5 January 2011 from http://www.iea.nl/timss2011.html.

Morrison, R. (2006). *A survey of survey flaws: false precision and reliance on the Internet.* Accessed on Nov 4, 2011 from http://www.lawdepartmentmanagementblog.com/law_department_management/2006/08/a_survey_of_sur.html

Organization for Economic Co-operation and Development (OECD, 2010). *PISA 2009 Results: What Students Know and Can Do -- Student performance in Reading, Mathematics, and Science, Volume 1*. Retrieved on 5 January 2010 from http://www.oecd.org/document/53/0,3746,en_32252351_46584327_46584821_1_1_1_1,00.html

Revere, P. (2009). False precision and the state of health and science reporting. *Effect Measure*. Retrieved on Nov 4, 2011 from http://scienceblogs.com/effectmeasure/2009/07/false_precision_and_the_state.php

Robinson, J. P. (2010). The Effects of Test Translation on Young English Learners' Mathematics Performance. *Educational Researcher*, *39* (8), 582-590.

Singapore Department of Statistics. (2010). *Census of Population 2010: Statistical Release 1, Demographic Characteristics, Education, Language and Religion*. Ministry of Trade and Industry, Republic of Singapore, Chart 1, p. 141.

Soh, K. C. (2008). Effect size: What does it do for educational action researchers? *North Star, 1(1)*, 63-70.

Soh, K. C. (2011a). University Rankings: How Serious Should We Take Them? *Journal of Higher Education Policy and Management.* Paper revised for publication.

Soh, K. C. (2011b). Statistically speaking, correctly. *North Star, 2*(2), 108-127.

Soh, K. C. (2011c). At the rear mirror and through the wind screen: teachers becoming teacher-researchers in Singapore schools. *New Horizons in Education, 59*(1), 12-24.

Thompson. B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools, 5*(2), 33-88.

Van der Vijer, F. & Hambleton, R. K. (1996). Translating tests: some practical guidelines. *European Psychologist*, 1, 89-99.

The author appreciates the perceptive comments of the reviewers on the manuscript and is responsible for any errors in this published version

Author

Dr. SOH Kay Cheng
He was a retired Senior Fellow of National Institute of Education in Singapore. He was on our journal's advisory board during 1995-96, helped to review numerous articles with short notice, and wrote occasionally for the journal (e.g. Soh, 2011c).
[sohkc@singnet.com.sg]

Funding source of the article: Self-financed

Received: 3.12.11 , accepted: 23.4.12, revised: 27.4.12.